On pointwise adaptive curve estimation based on inhomogeneous data

Stéphane Gaiffas

ON POINTWISE ADAPTIVE CURVE ESTIMATION BASED ON INHOMOGENEOUS DATA

STÉPHANE GAÏFFAS

Laboratoire de Probabilités et Modèles Aléatoires
Université Paris 7, 175 rue du Chevaleret, 75013 Paris
email: gaiffas@math.jussieu.fr

Abstract. We want to recover a signal based on noisy inhomogeneous data (the amount of data can vary strongly on the estimation domain). We model the data using nonparametric regression with random design, and we focus on the estimation of the regression at a fixed point x_0 with little, or much data. We propose a method which adapts both to the local amount of data (the design density is unknown) and to the local smoothness of the regression function. The procedure consists of a local polynomial estimator with a Lepski type data-driven bandwidth selector, see for instance Lepski et al. (1997). We assess this procedure in the minimax setup, over a class of function with local smoothness $s > 0$ of Hölder type. We quantify the amount of data at x_0 in terms of a local property on the design density called regular variation, which allows situations with strong variations in the concentration of the observations. Moreover, the optimality of the procedure is proved within this framework.

1. Introduction

1.1. **The model.** We observe n pairs of random variables $(X_i, Y_i) \in \mathbb{R} \times \mathbb{R}$ independent and identically distributed satisfying

$$Y_i = f(X_i) + \xi_i, \tag{1.1}$$

where $f : [0,1] \to \mathbb{R}$ is the unknown signal to be recovered, the variables (ξ_i) are centered Gaussian with variance σ^2 and independent of the design $X_1, \dots, X_n$. The variables X_i are distributed with respect to a density μ. We want to recover f at a fixed point x_0.

The classical way of considering the nonparametric regression model is to take deterministics $X_i = i/n$. In this model with an equispaced design, the observations are *homogeneously* distributed over the unit interval. If we take random X_i, we can model cases with *inhomogeneous* observations as the design distribution is "far" from the uniform law. In particular, in order to include situations with little or much data in the model, we allow the density μ to be *degenerate* (vanishing or exploding) at x_0. In this problem, we are interested in the adaptive estimation of f at x_0, both adaptive to the smoothness of f and to the inhomogeneity of the data.

1.2. **Motivations.** The adaptive estimation of the regression is a well-developed problem. Several adaptive procedures can be applied for the estimation of a signal with unknown smoothness: nonlinear wavelet estimation (thresholding), model selection, kernel estimation with a variable bandwidth (the Lepski method), and so on. Recent results dealing with the

Date: February 6, 2006.
2000 *Mathematics Subject Classification.* 62G05, 62G08.
Key words and phrases. adaptive estimation, inhomogeneous data, nonparametric regression, random design.

adaptive estimation of the regression function when the design is not equispaced or random include Antoniadis et al. (1997), Baraud (2002), Brown and Cai (1998), Wong and Zheng (2002), Maxim (2003), Delouille et al. (2004), Kerkyacharian and Picard (2004), among others.

Here, we focus on a slightly different problem: our aim is to recover the signal locally, based on data which can be eventually very inhomogeneous. More precisely, we want to be able to handle simultaneously situations where the observations are very concentrated at the estimation point, or conversely, very defficient, with the aim to illustrate the consequences of inhomogeneity on the accuracy of estimation within the theory. This problem is considered in Gaïffas (2004), where several minimax rates are computed under several types of behaviours for the design density. The estimator which is proposed therein is adaptive to the inhomogeneity of the data, but not smoothness-adaptive. Therefore, the results presented here extend our previous work from Gaïffas (2004), since we construct a procedure which is here both adaptive to the local smoothness, and to the distribution of the data.

1.3. **Organisation of the paper.** We construct the adaptive estimator in section 2, and we assess this estimator in section 3. First, we give an upper bound in theorem 1 which is stated conditionally on the design. Then, we propose in section 3.2 a way of quantifying the local inhomogeneity of the data with an appropriate assumption on the local behaviour of the design density. Under this assumption, we provide another upper bound in theorem 2. In section 4, we discuss the optimality of the estimator, and we prove in theorem 3 that the convergence rate from theorem 2 is optimal. We discuss some technical points in section 5 and we present numerical illustrations in section 6 for several datasets. Section 7 is devoted to the proofs and some well-known analytic facts are briefly recalled in appendix.

2. Construction of the adaptive procedure

The procedure described here is a local polynomial estimator with an adaptive data-driven selection of the bandwidth (the design density and the smoothness are both unknown). We need to introduce some notations. We define the empirical sample measure

$$\bar{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

where δ is a Dirac mass, and for an interval I such that $\bar{\mu}_n(I) > 0$, we introduce the pseudo-scalar product

$$\langle f\,,\, g\rangle_I := \frac{1}{\bar{\mu}_n(I)} \int_I f g \, d\bar{\mu}_n, \tag{2.1}$$

and $\|\cdot\|_I$ the corresponding pseudo-norm.

2.1. **Local polynomial estimation.** We fix $K \in \mathbb{N}$ and an interval I (typically $I = [x_0 - h, x_0 + h]$ where $h > 0$), which is a smoothing parameter that we call *bandwidth*. The idea is to look for the polynomial $\bar{f}_I$ of order K which is the closest to the data in the least square sense, with respect to the localised design-adapted norm $\|\cdot\|_I$:

$$\bar{f}_I := \operatorname*{argmin}_{g \in V_K} \|Y - g\|_I^2, \tag{2.2}$$

where V_K is the set of all real polynomials of order at most K. We can rewrite (2.2) in a variational form, in which we look for $\bar{f}_I \in V_K$ such that for any $\phi \in V_K$,

$$\langle \bar{f}_I\,,\, \phi\rangle_I = \langle Y\,,\, \phi\rangle_I, \tag{2.3}$$

where it suffices to consider only the power functions $\phi_p(\cdot) = (\cdot - x_0)^p$, $0 \leqslant p \leqslant K$. The coefficients vector $\bar{\theta}_I \in \mathbb{R}^{K+1}$ of the polynomial $\bar{f}_I$ is therefore solution, when it makes sense, of the linear system

$$\mathbf{X}_I \theta = \mathbf{Y}_I,$$

where for $0 \leqslant p, q \leqslant K$:

$$(\mathbf{X}_I)_{p,q} := \langle \phi_p \, , \, \phi_q \rangle_I \quad \text{and} \quad (\mathbf{Y}_I)_p := \langle Y \, , \, \phi_p \rangle_I. \tag{2.4}$$

The parameter $f(x_0)$ is then estimated by $\bar{f}_I(x_0)$. This linear method of estimation, called *local polynomial estimator* is well-known, see for instance Stone (1980), Fan and Gijbels (1995, 1996) and Tsybakov (2003) among many others.

In this paper, we work with a slightly modified version of the local polynomial estimator, which is convenient in situations with little or much data. We introduce

$$\bar{\mathbf{X}}_I := \mathbf{X}_I + \frac{1}{\sqrt{n\bar{\mu}_n(I)}} \mathbf{I}_{K+1} \mathbf{1}_{\Omega_I^c},$$

where $\mathbf{I}_{K+1}$ is the identity matrix in $\mathbb{R}^{K+1}$ and $\Omega_I := \{\lambda(\mathbf{X}_I) > (n\bar{\mu}_n(I))^{-1/2}\}$, $\lambda(M)$ standing for the smallest eigenvalue of a matrix M. Then, when $\bar{\mu}(I) > 0$, we consider the solution $\widehat{\theta}_I$ of the linear system

$$\bar{\mathbf{X}}_I \theta = \mathbf{Y}_I, \tag{2.5}$$

and denote by $\widehat{f}_I \in V_K$ the polynomial with coefficients $\widehat{\theta}_I$. When $\bar{\mu}_n(I) = 0$, we take simply $\widehat{f}_I := 0$.

Typically, the bandwidth I is given by a balance equation between the bias and the variance of $\widehat{f}_I$. Consequently, it depends on the local smoothness of f via the bias term. Therefore, an adaptive technique is required when the smoothness is unknown, which is the case in pratical situations.

2.2. **Adaptive bandwidth selection.** The adaptive procedure described here is based on the method introduced by Lepski (1990), see also Lepski et al. (1997), Lepski and Spokoiny (1997) and Spokoiny (1998). If a familly of linear estimators can be well sorted by their respective variances (this is the case with kernel estimators in the white noise model, see Lepski and Spokoiny (1997)), the Lepski procedure selects the largest bandwidth such that the corresponding estimator does not differ significantly from estimators with a smaller bandwidth. Following this principle, we propose a method which adapts to the unknown smoothness, and additionally to the original Lepski method, to the distribution of the data (the design density is unknown), in particular in cases with little of much data.

The idea of the adaptive procedure is the following: when $\widehat{f}_I$ is close to f (I is well-chosen), we have for any $J \subset I$, $\phi \in V_K$

$$\langle \widehat{f}_J - \widehat{f}_I \, , \, \phi \rangle_J = \langle Y - \widehat{f}_I \, , \, \phi \rangle_J \approx \langle Y - f \, , \, \phi \rangle_J = \langle \xi \, , \, \phi \rangle_J,$$

which is a noise term. Then, in order to remove noise, we select the largest I such that this noise term remains smaller than an appropriate threshold, for any $J \subset I$ and ϕ_p, $0 \leqslant p \leqslant K$. The bandwidth is selected in a fixed set of intervals $\mathcal{I}_n$ called *grid* (which is a tuning parameter of the procedure that we describe below) as follows:

$$\widehat{I}_n := \operatorname*{argmax}_{I \in \mathcal{I}_n} \big\{ \bar{\mu}_n(I) \text{ s.t. } \forall J \subset I,\, I \in \mathcal{I}_n,\, \forall 0 \leqslant m \leqslant K, \\ |\langle \widehat{f}_J - \widehat{f}_I \, , \, \phi_m \rangle_J| \leqslant \|\phi_m\|_J T_n(I,J) \big\},$$

where

$$T_n(I,J) := \sigma \big[D_{\mathcal{I}} (\varrho_n \bar{\mu}_n(I))^{-1/2} + D_p C_K \big(\log(n\bar{\mu}_n(I))/(n\bar{\mu}_n(J)) \big)^{1/2} \big], \tag{2.6}$$

with $C_K := 1+(K+1)^{1/2}$, $D_p := 4(p+1)^{1/2}$ and $D_{\mathcal{I}} \geqslant \sqrt{2}$. The estimator is then given by

$$\widehat{f}_n(x_0) := \widehat{f}_{\widehat{I}_n}(x_0). \tag{2.7}$$

An appropriate choice of grid (for practical purposes) is the following: first, we sort the (X_i, Y_i) into $(X_{(i)}, Y_{(i)})$ such that $X_{(i)} \leqslant X_{(i+1)}$. Then, we consider j such that $x_0 \in [X_{(j)}, X_{(j+1)}]$ (if necessary, we take $X_{(0)} = 0$ and $X_{(n+1)} = 1$), we choose some $a > 1$ and we introduce

$$\mathcal{I}_n := \bigcup_{p=0}^{[\log_a(j+1)]} \bigcup_{q=0}^{[\log_a(n-j)]} \big[X_{(j+1-[a^p])}, X_{(j+[a^q])}\big]. \tag{2.8}$$

This choice of grid is convenient for practical purposes, since its cardinal is $O((\log n)^2)$, which entails that the selection of the bandwidth in such a grid is fast at least for a parameter a not too close to 1.

The estimator $\widehat{f}_n(x_0)$ only depends on K and on the grid $\mathcal{I}_n$ (which are chosen by the statistician). It does not depend on the smoothness of f nor any assumption on μ. In this sense, this estimator is both smoothness-adaptive and design-adaptive. The procedure is proved to be adaptive (see section 3) for any smoothness parameter (of Hölder type) smaller than $K+1$. Therefore, we can understand K as a parameter of complexity of the procedure.

3. Assessment of the procedure: upper bounds

3.1. **Conditionally on the design.** When no assumption is made on the local behaviour of μ, we can work conditionally on the design. The procedure is assessed in the following way: first, we consider an ideal *oracle* interval given by

$$I_{n,f} := \operatorname*{argmax}_{I \subset [0,1],\, x_0 \in I} \Big\{ \bar{\mu}_n(I) \text{ s.t. } \operatorname{osc} f(I) \leqslant \sigma D_{\mathcal{I}} \big(\varrho_n \bar{\mu}_n(I)\big)^{-1/2} \Big\}, \tag{3.1}$$

where $\varrho_n := n/\log n$, $D_{\mathcal{I}} \geqslant \sqrt{2}$ and $\operatorname{osc} f(I)$ is the local oscillation of f in I, defined by

$$\operatorname{osc} f(I) := \inf_{P \in V_K} \sup_{y \in I} |f(y) - P(y)|, \tag{3.2}$$

where we recall that V_K is the set of all real polynomials with order at most K. The local oscillation is a common way of measuring the smoothness of a function.

The interval $I_{n,f}$, which is not necessarily unique, makes the balance between the bias and the $\log n$-penalised variance of $\widehat{f}_I$. Therefore, it can be understood as an *ideal adaptive bandwidth*, see Lepski and Spokoiny (1997) and Spokoiny (1998). The $\log n$ term in (3.1) is the *payment for adaptation*, see section 4.1. We use the word "oracle" since this interval depends on f directly. This oracle interval is used to define

$$R_{n,f} := \sigma \big(\varrho_n \bar{\mu}_n(I_{n,f})\big)^{-1/2}, \tag{3.3}$$

which is a random normalisation (it depends on the local amount of data) assessing the adaptive procedure in the next theorem. We introduce also

$$\bar{I}_{n,f} := \operatorname*{argmax}_{I \in \mathcal{I}_n} \Big\{ \bar{\mu}_n(I) \text{ s.t. } \operatorname{osc} f(I) \leqslant \sigma D_{\mathcal{I}} \big(\varrho_n \bar{\mu}_n(I)\big)^{-1/2} \Big\}, \tag{3.4}$$

which is an oracle interval in the grid, and we define the matrices

$$\mathbf{\Lambda}_I := \operatorname{diag}(\|\phi_0\|_I^{-1}, \dots, \|\phi_K\|_I^{-1}) \text{ and } \mathcal{G}_I := \mathbf{\Lambda}_I \bar{\mathbf{X}}_I \mathbf{\Lambda}_I. \tag{3.5}$$

We denote by $\mathfrak{L}_{n,f}$ the smallest eigenvalue of $\mathcal{G}_{\bar{I}_{n,f}}$, by $\mathfrak{X}_n$ the sigma-algebra generated by $X_1, \dots, X_n$ and by $\mathbb{E}^n_{f,\mu}$ the expectation with respect to the joint law $\mathbb{P}^n_{f,\mu}$ of the observations (1.1). We recall that $\Omega_I = \big\{\lambda(\mathbf{X}_I) > (n\bar{\mu}_n(I))^{-1/2}\big\}$, see section 2.1.

Theorem 1. *When* $\|f\|_\infty < +\infty$, *we have on* $\Omega_{\bar{I}_{n,f}} \cap \{n\bar{\mu}_n(I_{n,f}) \geqslant 2\}$ *for any* $p > 0$, $n \geqslant K+1$:

$$\mathbb{E}^n_{f,\mu}\big\{\big(R^{-1}_{n,f}|\widehat{f}_n(x_0) - f(x_0)|\big)^p|\mathfrak{X}_n\big\} \leqslant A\mathfrak{L}^{-p}_{n,f} + B(\|f\|_\infty \vee 1)^p,$$

where A *and* B *are constants depending respectively on* p, K, a *and* σ, p, K, a.

Note that the upper bound in theorem 1 is non-asymptotic since it holds for any $n \geqslant K+1$. Under an appropriate assumption on the design density (see the next section), we can see that the probability of $\Omega_{\bar{I}_{n,f}}$ is large and that $\mathfrak{L}_{n,f}$ is positive with a large probability. For more details, see section 5.

3.2. **How to quantify the local inhomogeneity of the data?** In this section, we propose a way of modeling situations where the amount of data is large or little at the estimation point x_0. The idea is simple: we allow the design density μ to be vanishing or exploding at x_0 with a power function behaviour type, which is quantified by a coefficient β called *index of regular variation*. Regular variation is a well-known notion, commonly used for quantifying the asymptotic behaviour of probability queues. It is also intimately linked with the theory of extreme values. On regular variation, we refer to Bingham et al. (1989).

Definition 1 (Regular variation)**.** A function $g : \mathbb{R}^+ \to \mathbb{R}^+$ is regularly varying at 0 if it is continuous, and if there is $\beta \in \mathbb{R}$ such that

$$\forall y > 0, \quad \lim_{h\to 0^+} g(yh)/g(h) = y^\beta. \tag{3.6}$$

We denote by $\mathrm{RV}(\beta)$ the set of all such functions. A function in $\mathrm{RV}(0)$ is *slowly varying*.

The assumption on μ is the following: we assume that there is $\nu > 0$ and $\beta > -1$ such that for any y, $|y - x_0| \leqslant \nu$:

$$\mu(x_0 + y) = \mu(x_0 - y) \ \text{ and } \ \mu(x_0 + \cdot\,) \in \mathrm{RV}(\beta). \tag{3.7}$$

This assumption means that μ is symmetrical within a neighbourhood of x_0, and varies regularly (on both sides). The local symmetry assumption is made in order to simplify the presentation of the results, but not necessary, see section 5 for more details. Note that this assumption includes the classical case with μ positive and continuous at x_0, and that in this case, $\beta = 0$.

3.3. **Minimax adaptive upper bound.** In this section, we assess the adaptive procedure $\widehat{f}_n$ in the minimax adaptive framework under assumption (3.7), which is an assumption quantifying the local amount the data. In what follows, we use the notation $I_h := [x_0 - h, x_0 + h]$. Let us consider the following smoothness class of function.

Definition 2. If $\omega \in \mathrm{RV}(s)$, $0 < s \leqslant K+1$ and $Q, \delta > 0$, we introduce

$$\mathcal{F}_\delta(\omega, Q) := \big\{f : \mathbb{R} \to \mathbb{R} \text{ s.t. } \|f\|_\infty \leqslant Q \text{ and } \forall h \leqslant \delta,\ \operatorname{osc} f(I_h) \leqslant \omega(h)\big\},$$

where we recall that $\operatorname{osc} f(I)$ is the local oscillation of f around x_0, see (3.2).

The parameter δ, which is taken small below (eventually going to 0 with n), is the length of the interval in which the smoothness assumption is made: this assumption is local. The parameter Q can be arbitrary large, but fixed. We need such a parameter since the next upper bound is stated uniformly over a collection of such classes. Note that the adaptive estimator does not depend on Q. The parameter ω measures the local smoothess. The assumption $\omega \in \mathrm{RV}(s)$ is convenient here for the computation of the convergence rate, and it is general since it includes for instance Hölder smoothness (take $\omega(h) = Lh^s$). Note that

s must be smaller than $K+1$, the degree of local polynomial smoothing of the adaptive estimator, which will be actually proved to be adaptive for any smoothess smaller than $K+1$.

Throughout what follows, we use the notation $\mu(I) := \int_I \mu(t)dt$. We introduce $h_n(\omega,\mu)$ as the smallest solution to

$$\omega(h) = \sigma\big(\varrho_n \mu(I_h)\big)^{-1/2}, \tag{3.8}$$

where we recall that $\varrho_n = n/\log n$. This quantity is well defined as n is large enough, since $h \mapsto \omega(h)^2\mu(I_h)$ is continuous and vanishing at 0. This equation is the deterministic counterpart (among symmetrical intervals) of the bias-variance equation (3.1). We introduce also

$$r_n(\omega,\mu) := \omega(h_n(\omega,\mu)), \tag{3.9}$$

which is proved to be the minimax adaptive convergence rate over the classes $\mathcal{F}_\delta(\omega,Q)$, see the next theorem and theorem 3 below.

Theorem 2. *We consider the adaptive estimator $\widehat{f}_n(x_0)$ defined by (2.7), with grid choice given by*

$$\mathcal{I}_n := \bigcup_{i=1}^n \Big\{ \big[x_0 - |X_i - x_0|, x_0 + |X_i - x_0|\big] \Big\} \tag{3.10}$$

instead of the grid (2.8) (we keep the same selection rule for the bandwidth). If μ satisfies (3.7), we have for any $\omega \in \mathrm{RV}(s)$ for $0 < s \leqslant K+1$, for any $p, Q > 0$ and any δ_n such that for some $\rho > 1$, $\delta_n \geqslant \rho h_n(\omega,\mu)$:

$$\mathrm{limsup}_n \sup_{f \in \mathcal{F}_{\delta_n}(\omega,Q)} \mathbb{E}^n_{f,\mu}\Big\{ \big(r_n(\omega,\mu)^{-1} |\widehat{f}_n(x_0) - f(x_0)|\big)^p \Big\} \leqslant C, \tag{3.11}$$

where C is a constant depending on p, K, Q, β. Moreover, we have

$$r_n(\omega,\mu) \sim P(\log n/n)^{s/(1+2s+\beta)} \ell_{\omega,\mu}(\log n/n) \text{ as } n \to +\infty, \tag{3.12}$$

where $a_n \sim b_n$ means $\lim_{n\to+\infty} a_n/b_n = 1$, where $\ell_{\omega,\mu}$ is a slowly varying function characterized by ω and μ, and where $P = P(s,\beta,\sigma) := \sigma^{2s/(1+2s+\beta)}$.

Remark. When $\omega(h) = Lh^s$ (Hölder smoothness) we have more precisely (for computations details, see the proof of lemma 6)

$$r_n(\omega,\mu) \sim P(\log n/n)^{s/(1+2s+\beta)} \ell_{\omega,\mu}(\log n/n) \text{ as } n \to +\infty,$$

where $P = P(s,\beta,\sigma,L) := \sigma^{2s/(1+2s+\beta)} L^{(\beta+1)/(1+2s+\beta)}$.

This theorem states an upper bound over the classes $\mathcal{F}_{\delta_n}(\omega,Q)$, which include Hölder classes. Note that the grid choice (3.10), which differs from the choice (2.8) used in theorem 1, is linked with the control of the $\mathfrak{X}_n$-measurable quantity $\mathfrak{L}_{n,f}$, related to the uniform control of the solution to the linear system (2.5). Additional comments about this theorem can be found in section 5.

3.4. **Explicit examples of rates.** Let $\beta > -1$, $s, L > 0$ and $\alpha, \gamma \in \mathbb{R}$. If $\int_0^h \mu(x_0+t)dt = h^{\beta+1}(\log(1/h))^\alpha$ for $0 \leqslant h \leqslant \nu$, and if $\omega(h) = Lh^s(\log(1/h))^\gamma$, then lemma 10 below and easy computations entail

$$r_n(\omega,\mu) \sim P\big(n(\log n)^{\alpha-1-\gamma(1+\beta)/s}\big)^{-s/(1+2s+\beta)}, \tag{3.13}$$

where $P = P(s,\beta,\sigma,L)$. This rate has to be compared with the minimax rate from Gaïffas (2004):

$$P\big(n(\log n)^{\alpha-\gamma(1+\beta)/s}\big)^{-s/(1+2s+\beta)},$$

where the only difference with (3.13) is α instead of $\alpha - 1$ in the logarithmic exponent. This loss, often called *payment for adaptation* in the literature, is unavoidable in view of theorem 3, see section 4 for more details.

In the usual case, namely when μ is positive and continuous at x_0, and when f is Hölder (that is, $\omega(h) = Lh^s$), we have $\alpha = \beta = \gamma = 0$ and we find back the usual pointwise minimax adaptive rate (see Lepski (1990), Brown and Low (1996)):

$$P(\log n/n)^{s/(1+2s)},$$

where $P = P(s, 0, \sigma, L) = \sigma^{2s/(1+2s)} L^{1/(1+2s)}$. For the same design but for the smoothness parameter $\omega(h) = Lh^s(\log(1/h))^{-s}$, we find the convergence rate

$$Pn^{-s/(1+2s)},$$

where $P = P(s, 0, \sigma, L)$ and where there is no log since we have more smoothness than in the s-Hölder case.

4. Minimax adaptive optimality of the estimator

4.1. **Payment for adaptation.** When μ satisfies (3.7) and if some $\omega \in \mathrm{RV}(s)$ is fixed, we know from Gaïffas (2004) that the minimax rate over $\mathcal{F}_\delta(\omega, Q)$ is equal to

$$n^{-s/(2s+1+\beta)} \ell_{\omega,\mu}(1/n), \tag{4.1}$$

where $\ell_{\omega,\mu}$ is a slowly varying function, characterized by ω and μ. In theorem 2, we proved that the adaptive method $\widehat{f}_n(x_0)$ converges with the rate

$$(\log n/n)^{s/(2s+1+\beta)} \ell_{\omega,\mu}(\log n/n), \tag{4.2}$$

which is slower than the minimax rate (4.1) because of the extra $\log n$ term. The aim of this section is to prove that this extra term in unavoidable.

In a model with homogeneous information (for instance white noise or regression with equidistant design), we know that adaptive estimation to the unknown smoothness without loss of efficiency is not possible for pointwise risks, even when the unknown signal belongs to one of two Hölder classes, see Lepski (1990), Brown and Low (1996) and Lepski and Spokoiny (1997). This means that local adaptation cannot be achieved for free: we have to pay an extra factor in the convergence rate, at least of order $(\log n)^{2s/(1+2s)}$ when estimating a function with Hölder smoothness s. The authors call this phenomenon *payment for adaptation*. Here, we intend to generalize this result to inhomogeneous data.

4.2. **A minimax adaptive lower bound.** First, let us introduce $H(s, L) := \mathcal{F}_\delta(\omega, Q)$ for $\omega(h) = Lh^s$, which is a classical local Hölder ball with smoothness s and radius L. Then, let $L' > L > 0$ and $s > s' > 0$ be such that $[s] = [s']$. We define $A := H(s', L')$ and $B := H(s, L)$. We denote by a_n (resp. b_n) the minimax rate given by (4.1) over A (resp. B) and by α_n (resp. β_n) the adaptive rate given by (4.2) over A (resp. B).

Theorem 3. *If an estimator $\widetilde{f}_n$ satisfies for some $p > 1$ the two following upper bounds (that is, it is asymptotically minimax over A and B):*

$$\mathrm{limsup}_n \sup_{f \in A} \mathbb{E}^n_{f,\mu}\left\{ \left(a_n^{-1} |\widetilde{f}_n(x_0) - f(x_0)|\right)^p \right\} < +\infty, \tag{4.3}$$

$$\mathrm{limsup}_n \sup_{f \in B} \mathbb{E}^n_{f,\mu}\left\{ \left(b_n^{-1} |\widetilde{f}_n(x_0) - f(x_0)|\right)^p \right\} < +\infty, \tag{4.4}$$

then:

$$\mathrm{liminf}_n \sup_{f \in A} \mathbb{E}^n_{f,\mu}\left\{ \left(\alpha_n^{-1} |\widetilde{f}_n(x_0) - f(x_0)|\right)^p \right\} > 0. \tag{4.5}$$

Note that (4.5) contradicts (4.3) since $\lim_n a_n/\alpha_n = 0$. The consequence is that *there is no pointwise minimax adaptive estimator over two such classes A and B and that the best achievable rate is α_n.*

5. Discussion

About theorem 1. Note that $\mathbf{X}_I = \mathbf{F}_I \mathbf{F}_I'$ where $\mathbf{F}_I$ is the matrix of size $n \times (K+1)$ with entries $(\mathbf{F}_I)_{i,m} = (X_i - x_0)^m$ for $0 \leqslant i \leqslant n$ and $0 \leqslant m \leqslant K$, and that $\ker \mathbf{X}_I = \ker \mathbf{F}_I$. This entails, when $n < K+1$, that $\mathbf{X}_I$ is not invertible since its kernel is not zero, and that Ω_I is empty. Therefore, theorem 1 is stated for $n \geqslant K+1$.

About assumption (3.7). Note that the local symmetry assumption is not necessary, but made in order to simplify the notations and the proofs. When there are $\beta^-, \beta^+ > -1$ such that $\mu(x_0 + \cdot) \in \mathrm{RV}(\beta^+)$ and $\mu(x_0 - \cdot) \in \mathrm{RV}(\beta^-)$ the result stated in theorem 2 is similar, where the convergence rate is again given by (3.12) with $\beta = \min(\beta^-, \beta^+)$, which means that the side with the largest amount of data "dominates" (asymptotically) the other one in the convergence rate.

About theorem 2. The reason why we need to take the grid (3.10) in theorem 2 is linked with the uniform control of $\mathfrak{L}_{n,f}$, which is necessary in the upper bound when solving the system (2.5). We can prove this theorem with the grid (2.8), which is more convenient in practice, if we assume that $\mathfrak{L}_{n,f} \geqslant \lambda$ for some $\lambda > 0$, uniformly for f in the union of $\mathcal{F}_\delta(\omega, Q)$ for $\omega \in \mathrm{RV}(s)$, $0 < s \leqslant K+1$. However, we have chosen to provide the upper bound only under assumption (3.7), which is used to quantify the local amount of data only.

Other remarks. The fact that the noise level σ is known is of little importance. If it is unknown we can plug-in some estimator $\widehat{\sigma}_n^2$ in place of σ^2. Following Gasser et al. (1986) or Buckley et al. (1988) we can consider for instance

$$\widehat{\sigma}_n^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{(i+1)} - Y_{(i)})^2, \tag{5.1}$$

where $Y_{(i)}$ is the observation at the point $X_{(i)}$ where $X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}$.

Literature. Bandwidth selection procedures in local polynomial estimation can be found in Fan and Gijbels (1995), Goldenshluger and Nemirovski (1997) or Spokoiny (1998). In this last paper the author is interested in the regression function estimation near a change point. The main idea and difference between the work by Spokoiny (1998) and the previous work by Goldenshluger and Nemirovski (1997) is to solve the linear problem (2.3) in a non-symmetrical neighbourhood I of x_0 not containing the change point. Our adaptive procedure is mainly inspired by the methods from Lepski and Spokoiny (1997), Lepski et al. (1997) and Spokoiny (1998), that we have modified in order to handle inhomogeneous data.

6. Simulated illustrations

In the simulations, we use the example signals from Donoho and Johnstone (1994). These functions are commonly used as benchmarks for adaptive estimators. We show in figure 1 the target functions and datasets with a uniform random design. The noise is Gaussian with σ chosen to have (root) signal-to-noise ratio 7. The sample size is $n = 2000$, which is large, but appropriate in order to exhibit the sensibility of the method to the underlying design.

We show the estimates in figure 2. For all estimates we consider the following tuning parameters of the procedure: the degree of the local polynomials is $K = 2$ and we consider the grid choice (2.8) with parameter $a = 1.05$. We recover the signal at each point $x = j/300$ with $j = 0, \ldots, 300$. The procedure is implemented in `C++` and is quite fast: it takes few seconds to recover the whole function at 300 points on a modern computer.

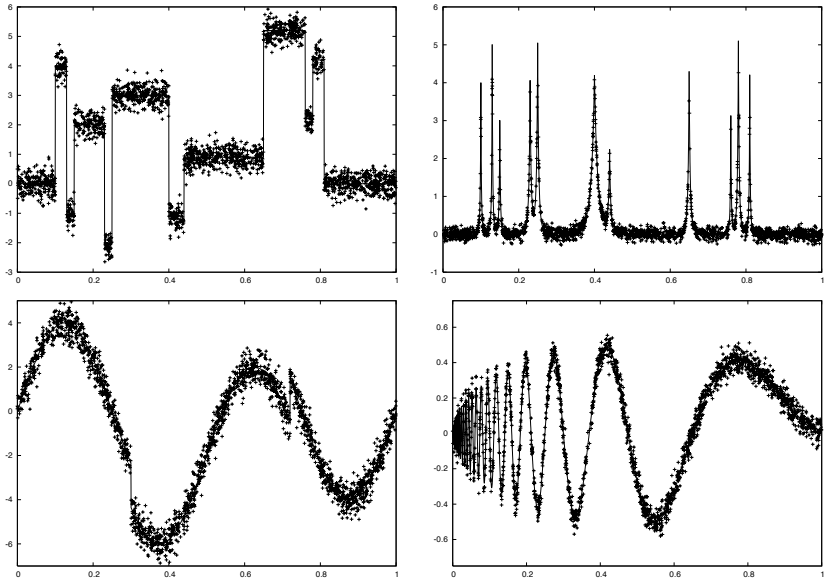

FIGURE 1. Blocks, bumps, heavysine and doppler with Gaussian noise and uniform design.

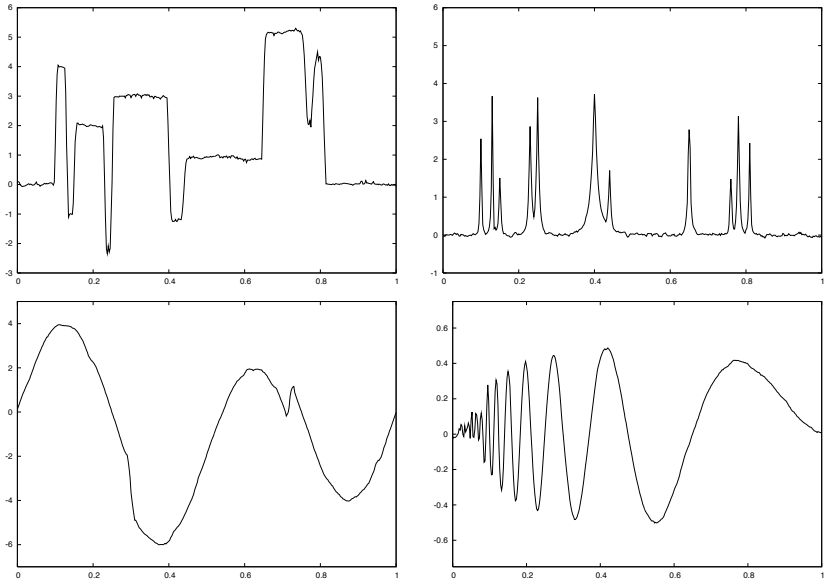

FIGURE 2. Estimates based on the datasets in figure 1.

Note that these estimates can be slightly improved with case by case tuned parameters: for instance, for the first dataset (blocks), the choice $K = 0$ gives a slightly better looking estimate (the target function is constant by parts). In figure 3 we show datasets with the same signal-to-noise ratio and sample size as in figure 1 but the design is non-uniform (we plot the design density on each of them). We show the estimates based on these datasets in figure 4. The same parameters as for figure 2 are used.

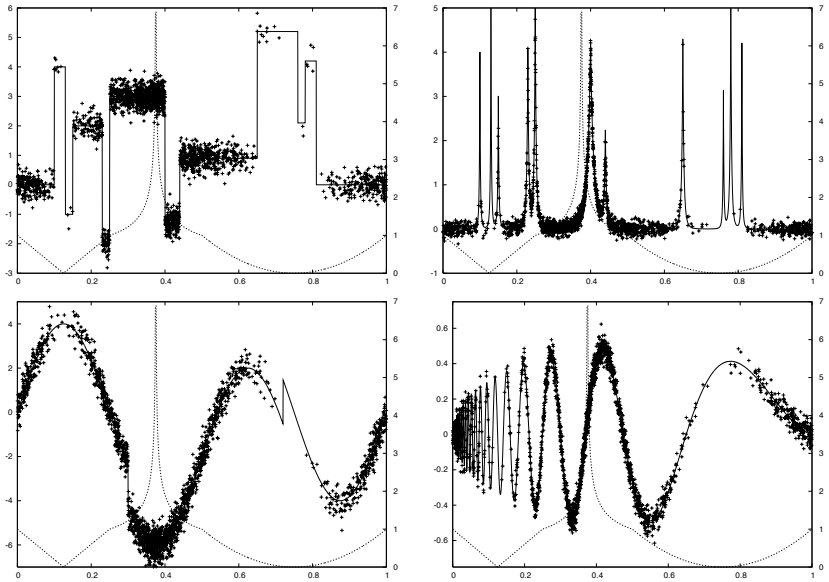

FIGURE 3. Blocks, bumps, heavysine and doppler with Gaussian noise and non-uniform design.

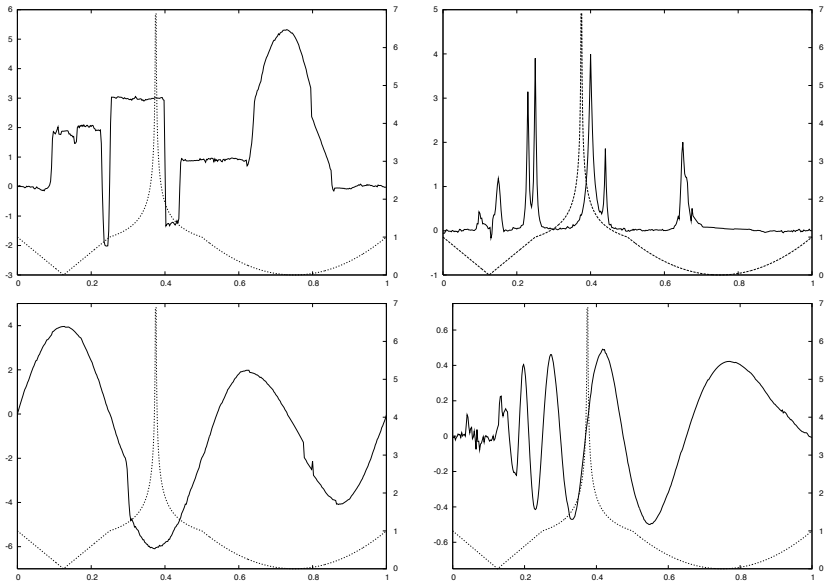

FIGURE 4. Estimates based on the datasets in figure 3.

In figures 5 and 6 we give a more localized illustration of the heavysine dataset. We keep the same signal-to-noise ratio and sample size. We consider the design density

$$\mu(x) = \frac{\beta+1}{x_0^{\beta+1} + (1-x_0)^{\beta+1}} \big|x - x_0\big|^{\beta} \mathbf{1}_{[0,1]}(x), \tag{6.1}$$

for $x_0 = 0.2, 0.72$ and $\beta = -0.5, 1$.

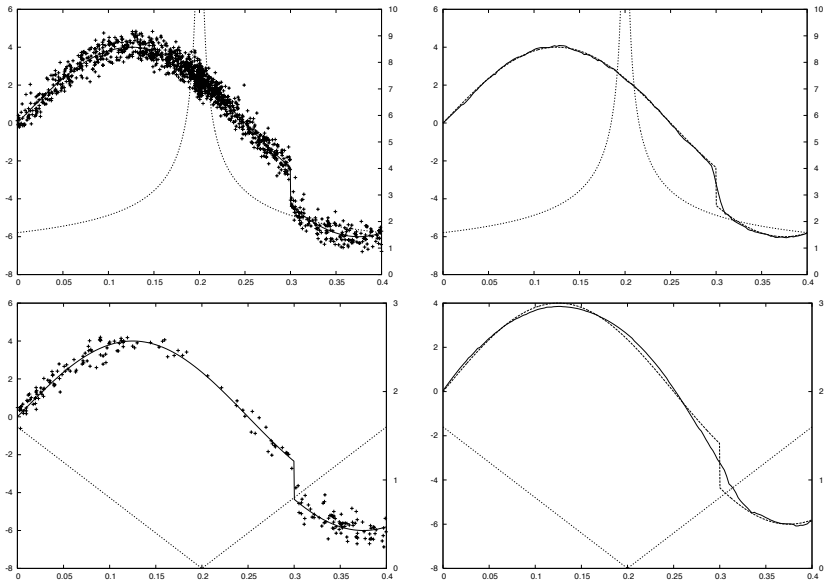

FIGURE 5. Heavysine datasets and estimates with design density (6.1) with $x_0 = 0.2$ and $\beta = -0.5$ at top, $\beta = 1$ at bottom.

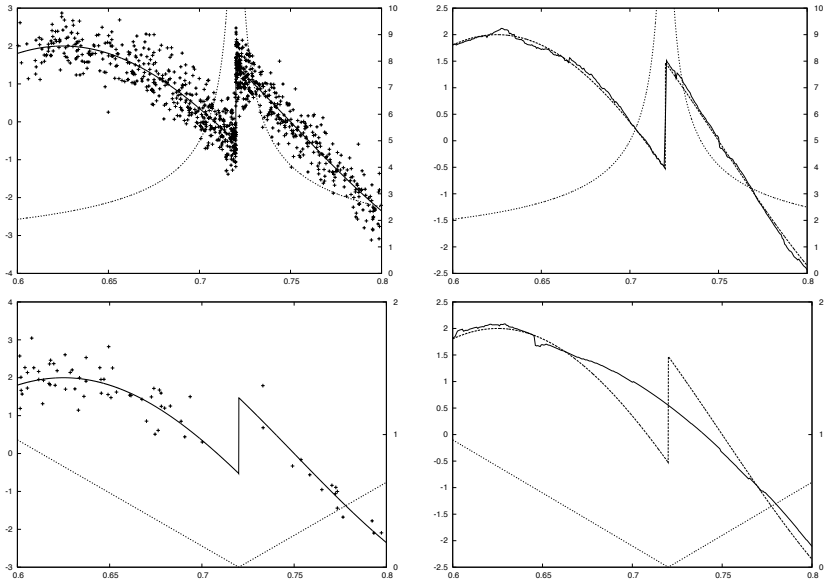

FIGURE 6. Heavysine datasets and estimates with design density (6.1) with $x_0 = 0.72$ and $\beta = -0.5$ at top, $\beta = 1$ at bottom.

7. Proofs

7.1. **Preparatory results and proof of theorem 1.** The next lemma is a version of the bias-variance decomposition of the local polynomial estimator, which is classical: see for instance Fan and Gijbels (1995, 1996), Goldenshluger and Nemirovski (1997), Spokoiny (1998) and Tsybakov (2003), among others. The next lemmas are used within the proof of theorem 1, and are proven below. We recall that $\Omega_I = \{\lambda(\mathbf{X}_I) > (n\bar{\mu}_n(I))^{-1/2}\}$, see section 2.1, that osc f stands for the local oscillation of f, see (3.2), and that the matrix $\mathcal{G}_I$ is defined in (3.5).

Lemma 1 (Bias variance decomposition)**.** *If I is such that $\bar{\mu}_n(I) > 0$ and $x_0 \in I$, we have on Ω_I that*

$$|\widehat{f}_I(x_0) - f(x_0)| \leqslant 2(K+1)^{1/2}\lambda(\mathcal{G}_I)^{-1}\big(\operatorname{osc} f(I) + \sigma(n\bar{\mu}_n(I))^{-1/2}|\gamma_I|\big), \tag{7.1}$$

where γ_I is, conditionally on $\mathfrak{X}_I$, centered Gaussian with $\mathbb{E}^n_{f,\mu}\{\gamma_I^2|\mathfrak{X}_n\} \leqslant 1$.

We introduce $m(p,\sigma) := (2/\pi)^{1/2}\int_{\mathbb{R}^+}(1+\sigma t)^p \exp(-t^2/2)dt$. The next lemma shows that the estimator cannot be too large in expectation.

Lemma 2. *When $\|f\|_\infty < +\infty$, we have for any $p > 0$ and $J \subset [0,1]$:*

$$\mathbb{E}^n_{f,\mu}\big\{|\widehat{f}_J(x_0)|^p|\mathfrak{X}_n\big\} \leqslant (K+1)^{p/2}m(p,\sigma)(\|f\|_\infty \vee 1)^p\big(n\bar{\mu}_n(J)\big)^{p/2}.$$

The next lemmas deal with the adaptive selection of the bandwidth. In particular, lemma 3 is of special importance, since it provides a control on the probability for a smoothing parameter to be selected by the procedure. Let us introduce

$$\mathcal{T}_{I,J,m} := \big\{|\langle \widehat{f}_J - \widehat{f}_I\,,\,\phi_m\rangle_J| \leqslant \sigma\|\phi_m\|_J T_n(I,J)\big\},$$

and $\mathcal{T}_{I,J} := \cap_{0\leqslant m\leqslant K}\mathcal{T}_{I,J,m}$, $\mathcal{T}_I := \cap_{J\in\mathcal{I}_n(I)}\mathcal{T}_{I,J}$, where $\mathcal{I}_n(I) := \{J \subset I \text{ s.t. } J \in \mathcal{I}_n\}$. Note that on $\mathcal{T}_I$, the bandwidth I is selected if it maximises $\bar{\mu}_n(I)$.

Lemma 3. *If $I \in \mathcal{I}_n$ is such that*

$$\operatorname{osc} f(I) \leqslant \sigma D_{\mathcal{I}}\big(\varrho_n\bar{\mu}_n(I)\big)^{-1/2},$$

where we recall that $D_{\mathcal{I}}$ is a tuning constant from the threshold term (2.6) and that $\varrho_n = n/\log n$, we have on $\Omega_I \cap \{n\bar{\mu}_n(I) \geqslant 2\}$:

$$\mathbb{P}^n_{f,\mu}\{\mathcal{T}_I^c|\mathfrak{X}_n\} \leqslant \#(\mathcal{I}_n(I))(K+1)(n\bar{\mu}_n(I))^{-D_p^2/8},$$

where we recall that $D_p = 4(p+1)^{1/2}$ (see (2.6)) and where $\#(E)$ denotes the cardinal of E.

Lemma 4. *Let $I \in \mathcal{I}_n$ and $J \in \mathcal{I}_n(I)$. On the event $\mathcal{T}_{I,J}\cap\Omega_J$, we have*

$$|\widehat{f}_I(x_0) - \widehat{f}_J(x_0)| \leqslant (K+1)^{1/2}\lambda(\mathcal{G}_J)^{-1}(D_{\mathcal{I}} + D_pC_K)\sigma\big(\varrho_n\bar{\mu}_n(J)\big)^{-1/2},$$

where we recall that $\mathcal{G}_J$ is given by (3.5).

Proof of theorem 1. Let j be such that $x_0 \in [X_{(j)}, X_{(j+1)}]$, where $X_{(i)} < X_{(i+1)}$ for any $1 \leqslant i \leqslant n$ (eventually, we take $X_{(0)} := 0$ and $X_{(n+1)} := 1$). We consider the largest interval $I^-_{n,f}$ in $\mathcal{I}_n$ such that $I^-_{n,f} \subset I_{n,f}$. Since osc $f(I)^2\bar{\mu}_n(I)$ increases as I increases, we have

$$\operatorname{osc} f(I^-_{n,f}) \leqslant \sigma D_{\mathcal{I}}\big(\varrho_n\bar{\mu}_n(I^-_{n,f})\big)^{-1/2},$$

thus $\bar{\mu}_n\big(I^-_{n,f}\big) \leqslant \bar{\mu}_n\big(\bar{I}_{n,f}\big)$. If p and q are such that

$$I^-_{n,f} = [X_{(j+1-[a^p])}, X_{(j+[a^q])}],$$

where $a > 1$ is the grid parameter (see (2.8)), and if u, v are such that $[X_{(u)}, X_{(v)}] \subset I_{n,f}$ and $\bar{\mu}_n([X_{(u)}, X_{(v)}]) = \bar{\mu}_n(I_{n,f})$, we have

$$\begin{aligned}\bar{\mu}_n\big([X_{(j+1-[a^p])}, X_{(j+[a^q])}]\big) &\leqslant \bar{\mu}_n\big([X_{(u)}, X_{(v)}]\big)\\ &\leqslant \bar{\mu}_n\big([X_{(j+1-[a^{p+1}])}, X_{(j+[a^{q+1}])}]\big),\end{aligned}$$

thus $\bar{\mu}_n\big(I_{n,f}\big) \leqslant a^2\bar{\mu}_n\big(I^-_{n,f}\big) \leqslant a^2\bar{\mu}_n\big(\bar{I}_{n,f}\big)$, and

$$\bar{\mu}_n\big(I_{n,f}\big)/a^2 \leqslant \bar{\mu}_n\big(\bar{I}_{n,f}\big) \leqslant \bar{\mu}_n\big(I_{n,f}\big). \tag{7.2}$$

Note that for the grid choice given by (2.8), we have

$$\#(\mathcal{I}_n(I)) \leqslant \big(\log(n\bar{\mu}_n(I))/\log a\big)^2. \tag{7.3}$$

We introduce $\mathrm{T}_n := \big\{\bar{\mu}_n(\bar{I}_{n,f}) \leqslant \bar{\mu}_n(\widehat{I}_n)\big\}$. By definition of $\widehat{I}_n$ we have $\mathrm{T}_n^c \subset \mathcal{T}^c_{\bar{I}_{n,f}}$. Using lemmas 2, 3 and (7.2), (7.3), we obtain:

$$\begin{aligned}&\mathbb{E}^n_{f,\mu}\big\{\big(R^{-1}_{n,f}|\widehat{f}_n(x_0) - f(x_0)|\big)^p \mathbf{1}_{\mathrm{T}_n^c}|\mathfrak{X}_n\big\}\\ &\quad\leqslant (2^{p-1}\vee 1)R^{-p}_{n,f}\Big(\big(\mathbb{E}^n_{f,\mu}\big\{|\widehat{f}_n(x_0)|^{2p}|\mathfrak{X}_n\big\}\big)^{1/2} + \|f\|^p_\infty\Big)\big(\mathbb{P}^n_{f,\mu}\{\mathcal{T}^c_{\bar{I}_{n,f}}|\mathfrak{X}_n\}\big)^{1/2}\\ &\quad\leqslant C(\sigma,a,p,K)(\|f\|_\infty\vee 1)^p \log\big(n\bar{\mu}_n(\bar{I}_{n,f})\big)^{1-p/2}\big(n\bar{\mu}_n(\bar{I}_{n,f}))\big)^{p-D_p^2/16}\\ &\quad\leqslant C(\sigma,a,p,K)(\|f\|_\infty\vee 1)^p,\end{aligned}$$

where $C(\sigma,a,p,K) := \sigma^{-p}(2^{p-1}\vee 1)(K+1)^{1+p/2}m(2p,\sigma)^{1/2}a^p/\log a$ and where we recall that $D_p = 4(p+1)^{1/2}$.

By the definition of $\widehat{I}_n$, we have

$$\mathrm{T}_n \subset \mathcal{T}_{\widehat{I}_n,\bar{I}_{n,f}},$$

thus using lemma 4 and (7.2), we obtain that on T_n,

$$|\widehat{f}_{\widehat{I}_n}(x_0) - \widehat{f}_{\bar{I}_{n,f}}(x_0)| \leqslant \lambda^{-1}(\mathcal{G}_{\bar{I}_{n,f}})(D_{\mathcal{I}} + D_pC_K)aR_{n,f}.$$

In view of lemma 1 and definition (3.4) of $\bar{I}_{n,f}$, we obtain using again (7.2):

$$\begin{aligned}|\widehat{f}_{\bar{I}_{n,f}}(x_0) - f(x_0)| &\leqslant \lambda(\mathcal{G}_{\bar{I}_{n,f}})^{-1}(K+1)^{1/2}\big(\operatorname{osc} f(\bar{I}_{n,f}) + \sigma(n\bar{\mu}_n(\bar{I}_{n,f}))^{-1/2}|\gamma_{\bar{I}_{n,f}}|\big)\\ &\leqslant \lambda(\mathcal{G}_{\bar{I}_{n,f}})^{-1}(K+1)^{1/2}\big(D_{\mathcal{I}} + (\log n)^{-1/2}|\gamma_{\bar{I}_{n,f}}|\big)aR_{n,f},\end{aligned}$$

where $\gamma_{\bar{I}_{n,f}}$ is, conditionally on $\mathfrak{X}_n$, centered Gaussian and such that $\mathbb{E}^n_{f,\mu}\{\gamma^2_{\bar{I}_{n,f}}|\mathfrak{X}_n\} \leqslant 1$. Then, we have on T_n

$$R^{-1}_{n,f}|\widehat{f}_n(x_0) - f(x_0)| \leqslant \lambda(\mathcal{G}_{\bar{I}_{n,f}})^{-1}(K+1)^{1/2}a\big(3D_{\mathcal{I}} + D_pC_K + (\log n)^{-1/2}|\gamma_{\bar{I}_n n,f}|\big),$$

and the theorem follows by integrating with respect to $\mathbb{P}^n_{f,\mu}(\cdot|\mathfrak{X}_n)$. □

7.2. **Preparatory results and proof of theorem 2.** Let us denote by $\mathbb{P}^n_\mu$ the joint probability of the variables X_i, $1 \leqslant i \leqslant n$ and let us recall the notation $\mu(I) = \int_I \mu(t)dt$. The next lemmas are used within the proof of theorem 2, their proofs can be found below.

Lemma 5. *For any $I \subset [0,1]$, $\varepsilon > 0$, we have*:

$$\mathbb{P}^n_\mu\big\{|\bar{\mu}_n(I)/\mu(I) - 1| > \varepsilon\big\} \leqslant 2\exp\Big(-\frac{\varepsilon^2}{1+\varepsilon/3}n\mu(I)\Big).$$

Lemma 6. *If μ satisfies (3.7), $\omega \in \mathrm{RV}(s)$ and $r_n = r_n(\omega, \mu)$ is given by (3.8) and (3.9), we have*

$$r_n \sim P(s, \beta, \sigma, 1)(\log n/n)^{s/(1+2s+\beta)} \ell_{\omega,\mu}(\log n/n) \text{ as } n \to +\infty, \tag{7.4}$$

where $\ell_{\omega,\mu} \in \mathrm{RV}(0)$ is characterised by ω and μ and where we recall that $P(s, \beta, \sigma, L) = \sigma^{2s/(1+2s+\beta)} L^{(\beta+1)/(1+2s+\beta)}$. When $\omega(h) = Lh^s$, $L > 0$ (Hölder smoothness), we have more precisely:

$$r_n \sim P(\sigma, \beta, \sigma, L)(\log n/n)^{s/(1+2s+\beta)} \ell_{\omega,\mu}(\log n/n) \text{ as } n \to +\infty. \tag{7.5}$$

We need to introduce some notations. If $\alpha \in \mathbb{N}$, $h > 0$ and if $\varepsilon > 0$, we define the event

$$\mathrm{D}_{n,\alpha}(\varepsilon, h) := \Big\{ \Big| \frac{1}{\mu(I_h)} \int_{I_h} \Big(\frac{\cdot - x_0}{h} \Big)^\alpha d\bar{\mu}_n - g_{\alpha,\beta} \Big| \leqslant \varepsilon \Big\},$$

where $g_{a,b} := (1 + (-1)^a)(b+1)/(2(a+b+1))$. The next lemmas are specifically linked with the uniform control of the smallest eigenvalue of the matrix $\mathbf{\Lambda}_I \bar{\mathbf{X}}_I \mathbf{\Lambda}_I$, defined by (3.5).

Lemma 7. *If μ satisfies (3.7), we have for any positive sequence (γ_n) going to 0 and any $\alpha \in \mathbb{N}, \varepsilon > 0$:*

$$\mathbb{P}_\mu^n \big\{ \mathrm{D}_{n,\alpha}(\varepsilon, \gamma_n)^c \big\} \leqslant 2 \exp \Big(- \frac{\varepsilon^2}{8(1 + \varepsilon/3)} n\mu(I_{\gamma_n}) \Big), \tag{7.6}$$

when n is large enough.

We recall that $h_n = h_n(\omega, \mu)$ is defined by (3.8). In what follows, we omit the dependence upon ω and μ to avoid overloaded notations. We introduce

$$H_n := \operatorname*{argmin}_{h \in [0,1]} \big\{ \omega(h) \geqslant \sigma(\varrho_n \bar{\mu}_n(I_h))^{-1/2} \big\}, \tag{7.7}$$

which is an approximation of h_n when μ is unknown. The next lemma controls the way how H_n and h_n are close. If $0 < \varepsilon < 1$, we introduce the event

$$\mathrm{C}_n(\varepsilon) := \{ (1 - \varepsilon) h_n < H_n \leqslant (1 + \varepsilon) h_n \}.$$

Lemma 8. *If $\omega \in \mathrm{RV}(s)$, $s > 0$ then for any $0 < \varepsilon_2 \leqslant 1/2$ there exists $0 < \varepsilon_3 \leqslant \varepsilon_2$ such that for n large enough*

$$\mathrm{D}_{n,0}(\varepsilon_3, (1 - \varepsilon_2) h_n) \cap \mathrm{D}_{n,0}(\varepsilon_3, (1 + \varepsilon_2) h_n) \subset \mathrm{C}_n(\varepsilon_2).$$

Let us denote $\mathcal{G}_n := \mathcal{G}_{I_{H_n}}$ and introduce the symmetrical matrix $\mathcal{G}$ with entries, for $0 \leqslant p, q \leqslant K$:

$$(\mathcal{G})_{p,q} := \frac{(1 + (-1)^{p+q})(2p + \beta + 1)^{1/2}(2q + \beta + 1)^{1/2}}{2(p + q + \beta + 1)}.$$

This matrix is the limit (in probability) of $\mathcal{G}_n$ as $n \to +\infty$. It is easy to see that $\lambda(\mathcal{G}) > 0$: note that $\mathcal{G} = \mathbf{\Lambda} \mathbf{X} \mathbf{\Lambda}$ where $\mathbf{\Lambda} = \mathrm{diag}\big[(1 + \beta)^{1/2}, (2 + \beta)^{1/2}, \ldots, (2K + 1 + \beta)^{1/2}\big]$, which is clearly invertible, and $\mathbf{X}$ has entries $(\mathbf{X})_{p,q} = (1 + (-1)^{p+q})/(2(p+q+\beta+1))$ for $0 \leqslant p, q \leqslant K$. Let us define the vector $\mathbf{p}(t) = (1, t, \ldots, t^k)$. Then, we have $\lambda(\mathbf{X}) > 0$: otherwise,

$$0 = \lambda(\mathbf{X}) = \langle \mathbf{x}\,, \mathbf{X}\mathbf{x} \rangle = \int_{-1}^{1} (\mathbf{x}' \mathbf{p}(t))^2 |t|^\beta dt,$$

where $\mathbf{x} \in \mathbb{R}^{K+1}$ is non-zero vector, which leads to a contradiction, since $t \mapsto \mathbf{x}' \mathbf{p}(t)$ is a polynomial ($\mathbf{x}'$ stands for the transposition of $\mathbf{x}$). Let us introduce the events $\mathrm{A}_n(\varepsilon) := \big\{ |\lambda(\mathcal{G}_n) - \lambda(\mathcal{G})| \leqslant \varepsilon \big\}$ for $\varepsilon > 0$ and for $\alpha \in \mathbb{N}$

$$\mathrm{B}_{n,\alpha}(\varepsilon) := \Big\{ \Big| \frac{1}{\mu(I_{h_n})} \int_{I_{H_n}} \Big(\frac{\cdot - x_0}{h_n} \Big)^\alpha d\bar{\mu}_n - g_{\alpha,\beta} \Big| \leqslant \varepsilon \Big\},$$

which differs from $\mathrm{D}_{n,\alpha}(\varepsilon, h_n)$ since the integral is taken over I_{H_n} instead of I_{h_n}.

Lemma 9. *If $\omega \in \mathrm{RV}(s)$, $s > 0$ and μ satisfies (3.7), we can find for any $0 < \varepsilon \leqslant 1/2$ an event $\mathcal{A}_n(\varepsilon) \in \mathfrak{X}_n$ such that*

$$\mathcal{A}_n(\varepsilon) \subset \mathrm{A}_n(\varepsilon) \cap \mathrm{B}_{n,0}(\varepsilon) \cap \mathrm{C}_n(\varepsilon) \tag{7.8}$$

for n large enough, and

$$\mathbb{P}_\mu^n\big\{\mathcal{A}_n(\varepsilon)^c\big\} \leqslant 4(K+2)\exp\big(-D_{\mathcal{A}}\, r_n^{-2}\big), \tag{7.9}$$

where $r_n := r_n(\omega, \mu)$ is given by (3.9) and $D_{\mathcal{A}} > 0$.

Proof of theorem 2. The proof of this theorem is based on the proof of theorem 1 and the previous lemmas. In the same fashion as in the proof of theorem 1, where we replace only equation (7.3) by

$$\#(\mathcal{I}_n(I)) \leqslant \big(n\bar{\mu}_n(I)\big)^2$$

since the grid choice is (3.10) instead of (2.8), we obtain that on $\Omega_{I_{H_n}} \cap \big\{n\bar{\mu}_n(I_{H_n}) \geqslant 2\big\}$:

$$\mathbb{E}_{f,\mu}^n\big\{\big(R_{n,f}^{-1}|\widehat{f}_n(x_0) - f(x_0)|\big)^p|\mathfrak{X}_n\big\} \leqslant A\lambda(\mathcal{G}_{I_{H_n}})^{-p} + B(\|f\|_\infty \vee 1)^p,$$

where we recall that H_n is defined by (7.7). Let us define $\varepsilon := \min(\rho - 1, \lambda/2)$ and consider the event $\mathcal{A}_n(\varepsilon)$ from lemma 9. On this event, since $\mathcal{A}_n(\varepsilon) \subset \mathrm{C}_n(\varepsilon)$, we have $\delta_n \geqslant (1+\varepsilon)h_n \geqslant H_n$, thus $f \in \mathcal{F}_\delta(\omega, Q)$ entails that we have either

$$\operatorname{osc} f(I_{H_n}) \leqslant \omega(H_n) = \sigma\big(\varrho_n\bar{\mu}_n(I_{H_n})\big)^{-1/2},$$

or

$$\operatorname{osc} f(I_{H_n}) \leqslant \omega(H_n) \leqslant \sigma\big((n\bar{\mu}_n(I_{H_n}) - 1)/\log n\big)^{-1/2},$$

which entails that in both cases $\operatorname{osc} f(I_{H_n}) \leqslant \sigma D_{\mathcal{I}}\big(\varrho_n\bar{\mu}_n(I_{H_n})\big)^{-1/2}$ since $D_{\mathcal{I}} \geqslant \sqrt{2}$, and that

$$\bar{\mu}_n(I_{H_n}) \leqslant \bar{\mu}_n(I_{n,f}). \tag{7.10}$$

Note that $\mathrm{B}_{n,0}(\varepsilon) = \big\{|\bar{\mu}_n(I_{H_n})/\mu(I_{h_n}) - 1| \leqslant \varepsilon\big\}$. Thus, on $\mathcal{A}_n(\varepsilon)$, we have in view of (3.8), (3.9), (7.8) and (7.10) that $r_n(\omega,\mu)^{-1} \leqslant (1-\varepsilon)^{-1}R_{n,f}^{-1}$, and $n\bar{\mu}_n(I_{n,f}) \geqslant (1-\varepsilon)n\mu(I_{h_n}) \to +\infty$ as $n \to +\infty$, thus $\mathcal{A}_n(\varepsilon) \subset \Omega_{I_{H_n}} \cap \big\{n\bar{\mu}_n(I_{H_n}) \geqslant 2\big\}$. Then, since $\mathcal{A}_n(\varepsilon) \subset \mathrm{A}_n(\varepsilon)$, we have uniformly for $f \in \mathcal{F}_\delta(\omega, Q)$:

$$\mathbb{E}_{f,\mu}^n\big\{\big(r_n(\omega,\mu)^{-1}|\widehat{f}_n(x_0) - f(x_0)|\big)^p\mathbf{1}_{\mathcal{A}_n(\varepsilon)}\big\} \leqslant (1-\varepsilon)^{-p/2}(A(\lambda-\varepsilon)^{-p} + B(Q\vee 1)^p).$$

Now we work on the complementary $\mathcal{A}_n(\varepsilon)^c$. Using lemma 2 and (7.9), we obtain that uniformly for $f \in \mathcal{F}_\delta(\omega, Q)$:

$$\begin{aligned}
&\mathbb{E}_{f,\mu}^n\big\{\big(r_n^{-1}|\widehat{f}_n(x_0) - f(x_0)|\big)^p\mathbf{1}_{\mathcal{A}_n(\varepsilon)^c}\big\} \\
&\quad\leqslant (2^{p-1}\vee 1)r_n^{-p}\Big[\big(\mathbb{E}_{f,\mu}^n\big\{|\widehat{f}_n(x_0)|^{2p}\big\}\big)^{1/2} + Q^p\Big]\big(\mathbb{P}_\mu^n\{\mathcal{A}_n(\varepsilon)^c\}\big)^{1/2} \\
&\quad\leqslant (2^{p-1}\vee 1)(Q\vee 1)^p(1 + (K+1)^{p/2}m(\sigma,2p)^{1/2})r_n^{-p}n^{p/2}\big(\mathbb{P}_\mu^n\{\mathcal{A}_n(\varepsilon)^c\}\big)^{1/2} = o_n(1),
\end{aligned}$$

which entails (3.11). Moreover, (3.12) follows from lemma 6, which concludes the proof. □

The next lemma is a technical tool for computing the explicits examples given in section 3.4. The proof can be found in Gaïffas (2004).

Lemma 10. *Let $a \in \mathbb{R}$ and $b > 0$. If $G(h) = h^b(\log(1/h))^a$, then we have*

$$G^{\leftarrow}(h) \sim b^{a/b}h^{1/b}(\log(1/h))^{-a/b} \text{ as } h \to 0^+.$$

7.3. **Proofs of the lemmas.** In the following, we denote by $\mathbf{P}_I$ the projection in the space V_K for the scalar product $\langle \cdot\,,\,\cdot \rangle_I$ which is given by (2.1). Note that on Ω_I, we have (see section (2.1))

$$\widehat{f}_I = \bar{f}_I = \mathbf{P}_I Y, \tag{7.11}$$

where $\mathbf{P}_I$ is the projection in V_K with respect to $\langle \cdot\,,\,\cdot \rangle_I$. We denote respectively by $\langle \cdot\,,\,\cdot \rangle$ and by $\|\cdot\|$ the Euclidean scalar product and the Euclidean norm in $\mathbb{R}^{K+1}$. We denote by $\|\cdot\|_\infty$ the sup norm in $\mathbb{R}^{K+1}$. We define $e_1 := (1, 0, \dots, 0) \in \mathbb{R}^{K+1}$.

Proof of lemma 1. On Ω_I, we have $\bar{\mathbf{X}}_I = \mathbf{X}_I$ and $\lambda(\mathbf{X}_I) > (n\bar{\mu}_n(I))^{-1/2} > 0$, thus $\mathbf{X}_I$ is invertible. Since $\mathbf{\Lambda}_I$ is clearly invertible on this event, $\mathcal{G}_I$ is also invertible. By definition of $\operatorname{osc} f(I)$, we can find a polynomial $P_I^\varepsilon \in V_K$ such that

$$\sup_{x \in I} |f(x) - P_I^\varepsilon(x)| \leqslant \operatorname{osc} f(I) + \varepsilon/\sqrt{n},$$

for any fixed $\varepsilon > 0$. If we denote by θ_I the coefficients vector of P_I^ε then

$$\begin{aligned} |\widehat{f}_I(x_0) - f(x_0)| &\leqslant |\langle \mathbf{\Lambda}_I^{-1}(\widehat{\theta}_I - \theta_I)\,,\, e_1 \rangle| + \operatorname{osc} f(I) + \varepsilon/\sqrt{n} \\ &= |\langle \mathcal{G}_I^{-1}\mathbf{\Lambda}_I \mathbf{X}_I(\widehat{\theta}_I - \theta_I)\,,\, e_1 \rangle| + \operatorname{osc} f(I) + \varepsilon/\sqrt{n}. \end{aligned}$$

In view of (2.3), we have on Ω_I for $m = 0, \dots, K$:

$$\begin{aligned} (\mathbf{X}_I(\widehat{\theta}_I - \theta_I))_m = \langle \widehat{f}_I - P_I^\varepsilon\,,\, \phi_m \rangle_I &= \langle Y - P_I^\varepsilon\,,\, \phi_m \rangle_I \\ &= \langle f - P_I^\varepsilon\,,\, \phi_m \rangle_I + \langle \xi\,,\, \phi_m \rangle_I =: B_{I,m} + V_{I,m}, \end{aligned}$$

thus the decomposition into bias and variance terms $\mathbf{X}_I(\widehat{\theta}_I - \theta_I) = B_I + V_I$, and

$$|\widehat{f}_I(x_0) - f(x_0)| \leqslant |\langle \mathcal{G}_I^{-1}\mathbf{\Lambda}_I B_I\,,\, e_1 \rangle| + |\langle \mathcal{G}_I^{-1}\mathbf{\Lambda}_I V_I\,,\, e_1 \rangle| + \operatorname{osc} f(I) + \varepsilon/\sqrt{n}.$$

We have

$$|\langle \mathcal{G}_I^{-1}\mathbf{\Lambda}_I B_I\,,\, e_1 \rangle| \leqslant \|\mathcal{G}_I^{-1}\mathbf{\Lambda}_I B_I\| \leqslant \|\mathcal{G}_I^{-1}\|\|\mathbf{\Lambda}_I B_I\| \leqslant \|\mathcal{G}_I^{-1}\|(K+1)^{1/2}\|\mathbf{\Lambda}_I B_I\|_\infty,$$

and for $0 \leqslant m \leqslant K$,

$$|(\mathbf{\Lambda}_I B_I)_m| = \|\phi_m\|^{-1}|\langle f - P_I^\varepsilon\,,\, \phi_m \rangle_I| \leqslant \|f - P_I^\varepsilon\|_I \leqslant \operatorname{osc} f(I) + \varepsilon/\sqrt{n}.$$

Since $\lambda(M)^{-1} = \|M^{-1}\|$ for any symmetrical and positive matrix M, and since $\|\mathbf{\Lambda}_I^{-1}\| \leqslant 1$, we have on Ω_I:

$$\|\mathcal{G}_I^{-1}\| = \|\mathbf{\Lambda}_I^{-1}\mathbf{X}_I^{-1}\mathbf{\Lambda}_I^{-1}\| \leqslant \|\mathbf{X}_I^{-1}\| = \lambda(\mathbf{X}_I)^{-1} \leqslant (n\bar{\mu}_n(I))^{1/2} \leqslant n^{1/2},$$

thus

$$|\langle \mathcal{G}_I^{-1}\mathbf{\Lambda}_I B_I\,,\, e_1 \rangle| \leqslant (K+1)^{1/2}\big(\|\mathcal{G}_I^{-1}\| \operatorname{osc} f(I) + \varepsilon\big).$$

Conditionally on $\mathfrak{X}_n$, the random vector V_I is centered Gaussian with covariance matrix $\sigma^2(n\bar{\mu}_n(I))^{-1}\mathbf{X}_I$. Thus $\mathcal{G}_I^{-1}\mathbf{\Lambda}_I V_I$ is again centered Gaussian, with covariance matrix

$$\sigma^2(n\bar{\mu}_n(I))^{-1}\mathcal{G}_I^{-1}\mathbf{\Lambda}_I\mathbf{X}_I\mathbf{\Lambda}_I\mathcal{G}_I^{-1} = \sigma^2(n\bar{\mu}_n(I))^{-1}\mathcal{G}_I^{-1},$$

and $\langle \mathcal{G}_I^{-1}\mathbf{\Lambda}_I V_I\,,\, e_1 \rangle$ is then centered Gaussian with variance

$$\sigma^2(n\bar{\mu}_n(I))^{-1}\langle e_1\,,\, \mathcal{G}_I^{-1}e_1 \rangle \leqslant \sigma^2(n\bar{\mu}_n(I))^{-1}\|\mathcal{G}_I^{-1}\|.$$

Since $\mathcal{G}_I$ is positive symmetrical and its entries are smaller than one in absolute value, $\|\mathcal{G}_I^{-1}\| = \lambda(\mathcal{G}_I)^{-1}$ and $\lambda(\mathcal{G}_I) = \inf_{\|x\|=1}\langle x\,,\, \mathcal{G}_I x \rangle \leqslant \|\mathcal{G}_I e_1\| \leqslant (K+1)^{1/2}$. Thus $\|\mathcal{G}_I^{-1}\| \leqslant (K+1)^{1/2}\|\mathcal{G}_I^{-1}\|^2$, and the lemma follows. □

Proof of lemma 2. If $\bar{\mu}_n(J) = 0$, we have $\widehat{f}_J = 0$ by definition and the result is obvious, thus we assume $\bar{\mu}_n(J) > 0$. Since $\lambda(\bar{\mathbf{X}}_J) \geqslant (n\bar{\mu}_n(J))^{-1/2} > 0$, $\bar{\mathbf{X}}_J$ and $\mathbf{\Lambda}_J$ are invertible and $\mathcal{G}_J$ also is. Thus,

$$\widehat{f}_J(x_0) = \langle \mathbf{\Lambda}_J^{-1}\widehat{\theta}_J\,,\, e_1\rangle = \langle \mathcal{G}_J^{-1}\mathbf{\Lambda}_J\bar{\mathbf{X}}_J\widehat{\theta}_J\,,\, e_1\rangle = \langle \mathcal{G}_J^{-1}\mathbf{\Lambda}_J\mathbf{Y}_J\,,\, e_1\rangle.$$

For any $0 \leqslant m \leqslant K$, we have

$$\begin{aligned}|(\mathbf{\Lambda}_J\mathbf{Y}_J)_m| &\leqslant \|\phi_m\|_J^{-1}\big(|\langle f\,,\,\phi_m\rangle_J| + |\langle \xi\,,\,\phi_m\rangle_J|\big)\\ &\leqslant \|f\|_J + \|\phi_m\|_J^{-1}|\langle \xi\,,\,\phi_m\rangle_J|\\ &\leqslant \|f\|_\infty + \|\phi_m\|_J^{-1}|\langle \xi\,,\,\phi_m\rangle_J| =: \|f\|_\infty + |V_{J,m}|.\end{aligned}$$

Conditionally on $\mathfrak{X}_n$, the vector V_J with entries $(V_{J,m}; 0 \leqslant m \leqslant K)$ is centered Gaussian with variance $\sigma^2\big(n\bar{\mu}_n(J)\big)^{-1}\mathbf{\Lambda}_J\mathbf{X}_J\mathbf{\Lambda}_J$, thus $\mathcal{G}_J^{-1}V_J$ is also centered Gaussian, with variance

$$\sigma^2\big(n\bar{\mu}_n(J)\big)^{-1}\mathcal{G}_J^{-1}\mathbf{\Lambda}_J\mathbf{X}_J\mathbf{\Lambda}_J\mathcal{G}_J^{-1} = \sigma^2\big(n\bar{\mu}_n(J)\big)^{-1}\mathbf{\Lambda}_J^{-1}\bar{\mathbf{X}}_J^{-1}\mathbf{X}_J\bar{\mathbf{X}}_J^{-1}\mathbf{\Lambda}_J^{-1}.$$

The variable $\langle \mathcal{G}_J^{-1}V_J\,,\, e_1\rangle$ is then, conditionally on $\mathfrak{X}_n$, centered Gaussian with variance

$$\sigma^2\big(n\bar{\mu}_n(J)\big)^{-1}\langle e_1\,,\, \mathbf{\Lambda}_J^{-1}\bar{\mathbf{X}}_J^{-1}\mathbf{X}_J\bar{\mathbf{X}}_J^{-1}\mathbf{\Lambda}_J^{-1}e_1\rangle \leqslant \sigma^2\big(n\bar{\mu}_n(J)\big)^{-1}\|\mathbf{\Lambda}_J^{-1}\|^2\|\bar{\mathbf{X}}_J^{-1}\|^2\|\mathbf{X}_J\|,$$

and since clearly $\|\mathbf{X}_J\| \leqslant K+1$, $\|\mathbf{\Lambda}_J^{-1}\| \leqslant 1$ and $\|\bar{\mathbf{X}}_J^{-1}\| = \lambda(\bar{\mathbf{X}}_J)^{-1} \leqslant \big(n\bar{\mu}_n(J)\big)^{1/2}$, the variance of $\langle \mathcal{G}_J^{-1}V_J\,,\, e_1\rangle$ is, conditionally on $\mathfrak{X}_n$, smaller than $\sigma^2(K+1)$. Moreover, $\|\mathcal{G}_J^{-1}\| \leqslant \|\mathbf{\Lambda}_J^{-1}\|\|\bar{\mathbf{X}}_J^{-1}\|\|\mathbf{\Lambda}_J^{-1}\| \leqslant \big(n\bar{\mu}_n(J)\big)^{1/2}$, thus

$$|\widehat{f}_J(x_0)| \leqslant (K+1)^{1/2}(\|f\|_\infty \vee 1)\big(n\bar{\mu}_n(J)\big)^{1/2}\big(1+\sigma|\gamma_J|\big),$$

where γ_J is, conditionally on $\mathfrak{X}_n$, centered Gaussian with variance smaller than 1. The lemma follows by integrating with respect to $\mathbb{P}^n_{f,\mu}(\cdot|\mathfrak{X}_n)$. □

Proof of lemma 3. Let $0 \leqslant m \leqslant K$ and $J \in \mathcal{I}_n(I)$. In view of (2.3) and (7.11), we have on Ω_I:

$$\begin{aligned}\langle \widehat{f}_J - \widehat{f}_I\,,\,\phi_m\rangle_J &= \langle Y - \widehat{f}_I\,,\,\phi_m\rangle_J\\ &= \langle f - \widehat{f}_I\,,\,\phi_m\rangle_J + \langle \xi\,,\,\phi_m\rangle_J\\ &= \langle f - \mathbf{P}_I f\,,\,\phi_m\rangle_J + \langle \mathbf{P}_I f - \widehat{f}_I\,,\,\phi_m\rangle_J + \langle \xi\,,\,\phi_m\rangle_J\\ &= \langle f - \mathbf{P}_I f\,,\,\phi_m\rangle_J + \langle \mathbf{P}_I(f-Y)\,,\,\phi_m\rangle_J + \langle \xi\,,\,\phi_m\rangle_J\\ &= \langle f - \mathbf{P}_I f\,,\,\phi_m\rangle_J - \langle \mathbf{P}_I\xi\,,\,\phi_m\rangle_J + \langle \xi\,,\,\phi_m\rangle_J\\ &:= A + B + C.\end{aligned}$$

The term A is a bias term. By the definition of $\operatorname{osc} f(I)$ we can find a polynomial $P_I^\varepsilon \in V_K$ such that

$$\sup_{x\in I}|f(x) - P_I^\varepsilon(x)| \leqslant \operatorname{osc} f(I) + \varepsilon_n,$$

where $\varepsilon_n := \sigma D_p C_K \log 2/(4n)$. Thus, since $J \subset I$, $P_I^\varepsilon \in V_K$ and $\mathbf{P}_I$ is an orthogonal projection with respect to $\langle\cdot\,,\,\cdot\rangle_I$,

$$\begin{aligned}|A| \leqslant \|f - \mathbf{P}_I f\|_J\|\phi_m\|_J &\leqslant \|f - P_I^\varepsilon - \mathbf{P}_I(f - P_I^\varepsilon)\|_I\|\phi_m\|_J\\ &\leqslant \|f - P_I^\varepsilon\|_I\|\phi_m\|_J \leqslant (\operatorname{osc} f(I) + \varepsilon_n)\|\phi_m\|_J,\end{aligned}$$

and by assumption,

$$|A| \leqslant \|\phi_m\|_J\big(\sigma D_{\mathcal{I}}(\varrho_n\bar{\mu}_n(I))^{-1/2} + \varepsilon_n\big). \tag{7.12}$$

Conditionally on $\mathfrak{X}_n$, B and C are centered Gaussian. Clearly, C is centered Gaussian with variance

$$\sigma^2\|\phi_m\|_J^2/(n\bar{\mu}_n(I)).$$

Since $\mathbf{P}_I\xi$ has covariance matrix $\sigma^2\mathbf{P}_I\mathbf{P}_I' = \sigma^2\mathbf{P}_I$ ($\mathbf{P}_I$ is an orthogonal projection), the variance of B is equal to

$$\begin{aligned}\mathbb{E}_{f,\mu}^n\big\{\langle\mathbf{P}_I\xi\,,\,\phi_m\rangle_J^2|\mathfrak{X}_n\big\} &\leqslant \|\phi_m\|_J^2\mathbb{E}_{f,\mu}^n\{\|\mathbf{P}_I\xi\|_J^2|\mathfrak{X}_n\}\\ &= \|\phi_m\|_J^2\,\mathrm{Tr}\left(\mathbb{V}\mathrm{ar}(\mathbf{P}_I\xi|\mathfrak{X}_n)\right)/(n\bar{\mu}_n(J))\\ &= \sigma^2\|\phi_m\|_J^2\,\mathrm{Tr}(\mathbf{P}_I)/(n\bar{\mu}_n(J)),\end{aligned}$$

where $\mathrm{Tr}(M)$ stands for the trace of a matrix M. Since $\mathbf{P}_I$ is the projection on V_K, it follows that $\mathrm{Tr}(\mathbf{P}_I)\leqslant K+1$, and that the variance of B is smaller than

$$\sigma^2\|\phi_m\|_J^2(K+1)/(n\bar{\mu}_n(J)).$$

Then,

$$\mathbb{E}_{f,\mu}^n\{(B+C)^2|\mathfrak{X}_n\}\leqslant\sigma^2\|\phi_m\|_J^2C_K^2/(n\bar{\mu}_n(J)), \tag{7.13}$$

where we recall that $C_K = 1+(K+1)^{1/2}$. Since $n\bar{\mu}_n(I)\geqslant 2$ by assumption, and $\bar{\mu}_n(J)\leqslant 1$, we have

$$\varepsilon_n\leqslant\sigma D_pC_K\big[\log\big(n\bar{\mu}_n(I)\big)/\big(4n\bar{\mu}_n(J)\big)\big]^{1/2}. \tag{7.14}$$

Then, equations (7.12), (7.14) and the definition of the threshold (2.6) together entail

$$\begin{aligned}&\big\{\|\phi_m\|_J^{-1}|\langle\widehat{f}_I-\widehat{f}_J\,,\,\phi_m\rangle_J|>T_n(I,J)\big\}\\ &\qquad\subset\Big\{\frac{\|\phi_m\|_J^{-1}|B+C|}{\sigma(n\bar{\mu}_n(J))^{-1/2}C_K}>D_p\big(\log(n\bar{\mu}_n(I))\big)^{1/2}/2\Big\}.\end{aligned}$$

Then, since

$$\mathcal{T}_{I,J}^c=\bigcup_{m=0}^K\big\{\|\phi_m\|_J^{-1}|\langle\widehat{f}_I-\widehat{f}_J\,,\,\phi_m\rangle_J|>T_n(I,J)\big\},$$

we obtain using (7.13) and the fact that $\mathbb{P}\{|N(0,1)|>x\}\leqslant\exp(-x^2/2)$:

$$\begin{aligned}\mathbb{P}_{f,\mu}^n\{\mathcal{T}_I^c|\mathfrak{X}_n\}&\leqslant\sum_{J\in\mathcal{I}_n(I)}\sum_{m=0}^K\exp(-D_p^2\log(n\bar{\mu}_n(I))/8)\\ &\leqslant\#(\mathcal{I}_n(I))(K+1)(n\bar{\mu}_n(I))^{-D_p^2/8},\end{aligned}$$

which concludes the lemma. □

Proof of lemma 4. Let us define $\mathbf{H}_J := \mathbf{\Lambda}_J\mathbf{X}_J$. On Ω_J, we have:

$$\begin{aligned}|\widehat{f}_I(x_0)-\widehat{f}_J(x_0)| = |(\widehat{\theta}_I-\widehat{\theta}_J)_0| &\leqslant\|\mathbf{\Lambda}_J^{-1}(\widehat{\theta}_I-\widehat{\theta}_J)\|_\infty\\ &\leqslant\|\mathcal{G}_J^{-1}\mathbf{H}_J(\widehat{\theta}_I-\widehat{\theta}_J)\|_\infty\\ &\leqslant(K+1)^{1/2}\lambda(\mathcal{G}_J)^{-1}\|\mathbf{H}_J(\widehat{\theta}_I-\widehat{\theta}_J)\|_\infty.\end{aligned}$$

Since on Ω_J, $\langle\widehat{f}_I-\widehat{f}_J\,,\,\phi_m\rangle_J/\|\phi_m\|_J = (\mathbf{H}_J(\widehat{\theta}_I-\widehat{\theta}_J))_m$, and since $J\subset I$, we obtain that on $\mathcal{T}_{I,J}$:

$$\begin{aligned}|\widehat{f}_I(x_0)-\widehat{f}_J(x_0)|&\leqslant(K+1)^{1/2}\lambda(\mathcal{G}_J)^{-1}T_n(I,J)\\ &\leqslant(K+1)^{1/2}\lambda(\mathcal{G}_J)^{-1}\sigma(D_{\mathcal{I}}+D_pC_K)\big(\varrho_n\bar{\mu}_n(J)\big)\big)^{-1/2},\end{aligned}$$

thus the lemma. □

Proof of lemma 5. It suffices to use the Bernstein inequality to the sum of independent random variables $Z_i = \mathbf{1}_{X_i \in I} - \mu(I)$ for $1 \leqslant i \leqslant n$. □

Proof of lemma 6. Let us define $G(h) := \omega^2(h)\mu(I_h)$. In view of (3.7), we have $\mu(I_h) = 2\int_0^h \mu(x_0+t)dt$ for $h \leqslant \nu$, and $\mu(I_h) \in \mathrm{RV}(\beta+1)$ since $\beta > -1$ (see appendix), thus $G \in \mathrm{RV}(1+2s+\beta)$. The function G is continuous and going to 0 as $h \to 0^+$ since $1+2s+\beta > 0$, see (A.2). Thus, we have for n large enough $h_n = G^{\leftarrow}(\sigma^2\varrho_n^{-1})$ where $G^{\leftarrow}(h) := \inf\{y \geqslant 0 | G(y) \geqslant h\}$ is the generalized inverse of G. Since $G^{\leftarrow} \in \mathrm{RV}(1/(1+2s+\beta))$ (see appendix) we have $\omega \circ G^{\leftarrow} \in \mathrm{RV}(s/(1+2s+\beta))$ and we can write $\omega \circ G^{\leftarrow}(h) = h^{s/(1+2s+\beta)}\ell_{\omega,\mu}(h)$ where $\ell_{\omega,\mu}$ is slowly varying. In particular, we have $\ell_{\omega,\mu}(\sigma^2\varrho_n^{-1}) \sim \ell_{\omega,\mu}(\varrho_n^{-1})$, thus

$$r_n = \omega \circ G^{\leftarrow}(\sigma^2\varrho_n^{-1}) \sim P(s,\beta,\sigma,1)(\log n/n)^{s/(1+2s+\beta)}\ell_{\omega,\mu}(\log n/n) \text{ as } n \to +\infty.$$

When $\omega(h) = Lh^s$ we can write more precisely $h_n = G^{\leftarrow}\big((\sigma/L)^2\varrho_n^{-1}\big)$ where $G(h) = h^{2s}\mu(I_h)$, and we obtain (7.5) in the same fashion as (7.4). □

Proof of lemma 7. Let us define $Q_i := \big(\frac{X_i - x_0}{\gamma_n}\big)^{\alpha}\mathbf{1}_{X_i \in I_{\gamma_n}}$ and $Z_i := Q_i - \mathbb{E}_\mu^n\{Q_i\}$. In view of (3.7), we can find N such that $\gamma_n \leqslant \nu$ for any $n \geqslant N$ and:

$$\frac{1}{\mu(I_{\gamma_n})}\mathbb{E}_\mu^n\{Q_i\} = \frac{1+(-1)^\alpha}{2}\frac{\gamma_n^{\beta+1}\ell_\mu(\gamma_n)}{\int_0^{\gamma_n} t^\beta \ell_\mu(t)dt}\frac{\int_0^{\gamma_n} t^{\alpha+\beta}\ell_\mu(t)dt}{\gamma_n^{\alpha+\beta+1}\ell_\mu(\gamma_n)},$$

where for $h \leqslant \nu$, $\ell_\mu(h) = h^{-\beta}\mu(x_0+h) = h^{-\beta}\mu(x_0-h)$ is slowly varying (see appendix). Then, we have in view of (A.3):

$$\lim_{n\to+\infty}\frac{1}{\mu(I_{\gamma_n})}\mathbb{E}_\mu^n\{Q_i\} = g_{\alpha,\beta},$$

which entails that for n large enough:

$$\mathrm{D}_{n,\alpha}(\varepsilon,\gamma_n)^c \subset \Big\{\frac{1}{n\mu(I_{\gamma_n})}\Big|\sum_{i=1}^n Z_i\Big| > \varepsilon/2\Big\}. \tag{7.15}$$

Not that $\mathbb{E}_\mu^n\{Z_i\} = 0$, $|Z_i| \leqslant 2$, $\sum_{i=1}^n \mathbb{E}_\mu^n\{Z_i^2\} \leqslant n\mathbb{E}_\mu^n\{Q_i^2\} \leqslant n\mu(I_{\gamma_n})$ and that the Z_i, $1 \leqslant i \leqslant n$ are independent. Thus, we apply Bernstein inequality to the sum of the Z_i and the lemma follows. □

Proof of lemma 8. In view of (7.7), we have

$$\big\{H_n \leqslant (1+\varepsilon_2)h_n\big\} = \big\{\varrho_n\bar{\mu}_n(I_{(1+\varepsilon_2)h_n}) \geqslant \sigma^2\omega((1+\varepsilon_2)h_n)^{-2}\big\}.$$

We introduce $\varepsilon_3 := \min\big[\varepsilon_2, 1-(1-\varepsilon_2^2)^{-2}(1+\varepsilon_2)^{-2s}\big]$, which is positive for ε_2 small enough, and $\ell_\omega(h) := h^{-s}\omega(h)$ which is slowly varying, since $\omega \in \mathrm{RV}(s)$. Since (A.1) holds uniformly over each compact set in $(0,+\infty)$, we have for any $y \in [1/2, 3/2]$

$$(1-\varepsilon_2^2)\ell_\omega(h_n) \leqslant \ell_\omega(yh_n) \leqslant (1+\varepsilon_2^2)\ell_\omega(h_n) \tag{7.16}$$

for n large enough, so (7.16) with $y = 1+\varepsilon$ ($\varepsilon \leqslant 1/2$) entails in view of (3.8) and since $h \mapsto \mu(I_h)$ is increasing:

$$\begin{aligned}(1-\varepsilon_3)\varrho_n\mu(I_{(1+\varepsilon_2)h_n}) &\geqslant (1-\varepsilon_2^2)^{-2}(1+\varepsilon_2)^{-2s}\sigma^2\omega(h_n)^{-2}\\ &= \sigma^2\big((1+\varepsilon_2)h_n\big)^{-2s}(1-\varepsilon_2^2)^{-2}\ell_\omega(h_n)^{-2}\\ &\geqslant \sigma^2\omega((1+\varepsilon_2)h_n)^{-2}.\end{aligned}$$

Thus

$$\big\{\bar{\mu}_n(I_{(1+\varepsilon_2)h_n}) \geqslant (1-\varepsilon_3)\mu((1+\varepsilon_2)h_n)\big\} \subset \big\{H_n \leqslant (1+\varepsilon_2)h_n\big\},$$

and similarly on the other side, we have for n large enough:

$$\big\{\bar{\mu}_n(I_{(1-\varepsilon_2)h_n}) \leqslant (1+\varepsilon_3)\mu((1-\varepsilon_2)h_n)\big\} \subset \{(1-\varepsilon_2)h_n < H_n\},$$

thus the lemma. □

Proof of lemma 9. Since $\mathcal{G}_n$ and $\mathcal{G}$ are symmetrical, we get

$$\bigcap_{0\leqslant p,q\leqslant K} \big\{\big|(\mathcal{G}_n - \mathcal{G})_{p,q}\big| \leqslant \varepsilon/(K+1)^2\big\} \subset \mathrm{A}_n(\varepsilon),$$

where we used the fact that $\lambda(M) = \inf_{\|x\|=1}\langle x\,,\, Mx\rangle$ for any symmetrical matrix M. Then, an easy computation shows that if $\varepsilon_1 := \min\big[\varepsilon, \varepsilon(\beta+1)/((K+1)^2(2K+\beta+1))\big]$, we have for any $0 \leqslant p, q \leqslant K$:

$$\mathrm{B}_{n,p+q}(\varepsilon_1) \cap \mathrm{B}_{n,2p}(\varepsilon_1) \cap \mathrm{B}_{n,2q}(\varepsilon_1) \subset \big\{\big|(\mathcal{G}_n - \mathcal{G})_{p,q}\big| \leqslant \varepsilon/(K+1)^2\big\},$$

and then

$$\bigcap_{\alpha=0}^{2K} \mathrm{B}_{n,\alpha}(\varepsilon_1) \subset \mathrm{A}_n(\varepsilon).$$

Let us define $\varepsilon_2 := \varepsilon_1/(2(1+\varepsilon_1)^{2K+1})$ and let ε_3 be such that $(1+\varepsilon_3)^{\beta+3}/(1-\varepsilon_3) \leqslant 1+\varepsilon_2$ and $0 < \varepsilon_3 \leqslant \varepsilon_2$. Since $h \mapsto \bar{\mu}_n(I_h)$ is increasing, we have

$$\mathrm{C}_n(\varepsilon_3) \subset \big\{\bar{\mu}_n(I_{(1-\varepsilon_3)h_n}) \leqslant \bar{\mu}_n(I_{H_n}) \leqslant \bar{\mu}_n(I_{(1+\varepsilon_3)h_n})\big\},$$

and using lemma 8, we can find $0 < \varepsilon_4 \leqslant \varepsilon_3$ such that

$$\mathrm{D}_{n,0}(\varepsilon_4, (1-\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_4, (1+\varepsilon_3)h_n) \subset \mathrm{C}_n(\varepsilon_3).$$

In view of (A.1) and since $\ell_\mu(h) := h^{-(\beta+1)}\mu(I_h)$ is slowly varying, we have for any $0 < \varepsilon_3 \leqslant 1/2$:

$$\ell_\mu((1+\varepsilon_3)h_n) \leqslant (1+\varepsilon_3)\ell_\mu(h_n) \text{ and } \ell_\mu((1-\varepsilon_3)h_n) \geqslant (1-\varepsilon_3)\ell_\mu(h_n) \tag{7.17}$$

as n is large enough, thus simple algebra and the previous embeddings entail

$$\mathrm{D}_{n,0}(\varepsilon_4, (1-\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_4, (1+\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_3, h_n) \subset \Big\{\Big|\frac{\bar{\mu}_n(I_{H_n})}{\bar{\mu}_n(I_{h_n})} - 1\Big| \leqslant \varepsilon_2\Big\}.$$

Thus, in view of the previous embeddings, we have on $\mathrm{D}_{n,0}(\varepsilon_4, (1-\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_4, (1+\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_3, h_n)$:

$$\begin{aligned}
\frac{1}{\mu(I_{h_n})}\Big|\int_{I_{H_n}} \Big(\frac{\cdot - x_0}{h_n}\Big)^\alpha d\bar{\mu}_n - \int_{I_{h_n}} \Big(\frac{\cdot - x_0}{h_n}\Big)^\alpha d\bar{\mu}_n\Big| & \\
\leqslant \Big(\frac{H_n \vee h_n}{h_n}\Big)^\alpha \frac{\bar{\mu}_n(I_{h_n})}{\mu(I_{h_n})}\Big|\frac{\bar{\mu}_n(I_{H_n})}{\bar{\mu}_n(I_{h_n})} - 1\Big| & \\
\leqslant (1+\varepsilon_3)^\alpha(1+\varepsilon_3)\varepsilon_2 \leqslant (1+\varepsilon_1)^{2K+1}\varepsilon_2 = \varepsilon_1/2. &
\end{aligned}$$

Putting all the previous embeddings together, we obtain

$$\begin{aligned}
\mathrm{D}_{n,0}(\varepsilon_4, (1-\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_4, (1+\varepsilon_3)h_n) & \\
\cap\, \mathrm{D}_{n,0}(\varepsilon_4, h_n) \cap \mathrm{D}_{n,\alpha}(\varepsilon_1/2, h_n) \subset \mathrm{B}_{n,\alpha}(\varepsilon_1), &
\end{aligned}$$

and finally, equation (7.8) follows if we take

$$\begin{aligned}
\mathcal{A}_n(\varepsilon) := \mathrm{D}_{n,0}(\varepsilon_4, (1-\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_4, (1+\varepsilon_3)h_n) \cap \mathrm{D}_{n,0}(\varepsilon_4, h_n) & \\
\cap \bigcap_{0\leqslant\alpha\leqslant 2K} \mathrm{D}_{n,\alpha}(\varepsilon_4, h_n). &
\end{aligned}$$

In view of (3.8), (3.9), (7.17), and since $0 < \varepsilon_3 \leqslant 1/2$, we obtain using lemma 7:

$$\mathbb{P}_\mu^n\{\mathcal{A}_n(\varepsilon)^c\} \leqslant 4(K+2)\exp\big(-D_{\mathcal{A}}\, r_n^{-2}\big)$$

for n large enough, where $D_{\mathcal{A}} := 2^{-(\beta+2)}(\sigma\varepsilon_4)^2/(4(1+\varepsilon_4/3))$, which concludes the proof of the lemma. □

7.4. **Preparatory results and proof of theorem 3.** The proof of theorem 3 is similar to the proof of theorem 3 in Brown and Low (1996). It is based on the next theorem which can be found in Cai et al. (2004). This result is a general constrained risk inequality which is useful for several statistical problems, for instance superefficiency, adaptation and so on.

Let $p > 1$ and q be such that $1/p + 1/q = 1$ and X be a real random variable having distribution $\mathbb{P}_\theta$ with density f_θ. The parameter θ can take two values θ_1 or θ_2. We want to estimate θ based on X. The risk of an estimator δ based on X is given by

$$R_p(\delta,\theta) := \mathbb{E}_\theta\{|\delta(X)-\theta|^p\}.$$

We define $s(x) := f_{\theta_2}(x)/f_{\theta_1}(x)$ and $\Delta := |\theta_2 - \theta_1|$. Let

$$I_q = I_q(\theta_1,\theta_2) := \big(\mathbb{E}_{\theta_1}\{s^q(X)\}\big)^{1/q}.$$

Theorem 4 (Cai, Low and Zhao (2004))**.** *If δ is such that $R_p(\delta,\theta_1) \leqslant \varepsilon^p$ and if $\Delta > \varepsilon I_q$, we have*:

$$R_p(\delta,\theta_2) \geqslant (\Delta - \varepsilon I_q)^p \geqslant \Delta^p\Big(1 - \frac{p\varepsilon I_q}{\Delta}\Big).$$

The next proposition is a generalization of a result by Brown and Low (1996) for the random design model, when the data is inhomogeneous. Of course, in the classical case with μ continuous at x_0 and such that $\mu(x_0) > 0$, the result is barely the same as in Brown and Low (1996) with the same rates. This proposition is a lower bound for a superefficient estimator which implies directly the adaptive lower bound stated in theorem 3. Let us recall that a_n is the minimax rate and α_n is the minimax adaptive rate over A, see section 4.2.

Proposition 1. *If an estimator $\widetilde{f}_n$ based on (1.1) is asymptotically minimax over A, that is*:

$$\operatorname{limsup}_n \sup_{f\in A} \mathbb{E}_{f,\mu}^n\big\{\big(a_n^{-1}|\widetilde{f}_n(x_0) - f(x_0)|\big)^p\big\} < +\infty,$$

and if this estimator is superefficient at a function $f_0 \in A$, in the sense that for some $\gamma > 0$:

$$\operatorname{limsup}_n \mathbb{E}_{f_0,\mu}^n\big\{\big(a_n^{-1}n^\gamma|\widetilde{f}_n(x_0) - f_0(x_0)|\big)^p\big\} < +\infty, \tag{7.18}$$

then we can find another function $f_1 \in A$ such that

$$\operatorname{liminf}_n \inf_{\widetilde{f}_n} \mathbb{E}_{f_1,\mu}^n\big\{\big(\alpha_n^{-1}|\widetilde{f}_n(x_0) - f_1(x_0)|\big)^p\big\} > 0.$$

Proof of proposition 1. Since $\operatorname{limsup}_n \mathbb{E}_{f_0,\mu}\big\{\big(a_n^{-1}n^\gamma|\widetilde{f}_n(x_0) - f_0(x_0)|\big)^p\big\} = C < +\infty$, there is N such that for any $n \geqslant N$:

$$\mathbb{E}_{f_0,\mu}\big\{\big(|\widetilde{f}_n(x_0) - f_0(x_0)|\big)^p\big\} \leqslant 2Ca_n^p n^{-\gamma p}.$$

Let $k' = \lfloor s' \rfloor$ be the largest integer smaller than s'. Let g be k' times differentiable with support included in $[-1,1]$, $g(0) > 0$ and such that for any $|x| \leqslant \delta$, $|g^{(k')}(x) - g^{(k')}(0)| \leqslant k'!|x|^{s'-k'}$. Such a function clearly exists. We define

$$f_1(x) := f_0(x) + L'\rho_n^{s'} g\Big(\frac{x-x_0}{\rho_n}\Big),$$

where ρ_n is the smallest solution to

$$L' h^{s'} = \sigma\big(\varrho_n \mu(I_h)/b\big)^{-1/2},$$

where $b = 2g_\infty^{-2}(p-1)\gamma$, $g_\infty := \sup_x |g(x)|$ and where we recall that $\varrho_n = n/\log n$. We clearly have $f_1 \in A$. Let $\mathbb{P}_0^n, \mathbb{P}_1^n$ be the joint laws of the observations (1.1) when respectively $f = f_0$, $f = f_1$. A sufficient statistic for $\{\mathbb{P}_0^n, \mathbb{P}_1^n\}$ is given by $T_n := \log\big(\mathrm{d}\mathbb{P}_0^n/\mathrm{d}\mathbb{P}_1^n\big)$, and

$$T_n \sim \begin{cases} N\big(-\dfrac{v_n}{2}, v_n\big) & \text{under } \mathbb{P}_0^n, \\ N\big(\dfrac{v_n}{2}, v_n\big) & \text{under } \mathbb{P}_1^n, \end{cases}$$

where, by definition of ρ_n:

$$\begin{aligned} v_n = \frac{n}{\sigma^2}\|f_0 - f_1\|_{L^2(\mu)}^2 &= \frac{n}{\sigma^2}\int (f_0(x) - f_1(x))^2 \mu(x)dx \\ &\leqslant nL'^2 \rho_n^{2s'} \mu(I_{\rho_n}) g_\infty^2/\sigma^2 = 2(p-1)\gamma \log n. \end{aligned}$$

An easy computation gives $I_q = \exp\big(v_n(q-1)/2\big) \leqslant n^\gamma$, thus taking $\delta = \widehat{f}_n(x_0)$, $\theta_2 = f_1(x_0)$, $\theta_1 = f_0(x_0)$ and $\varepsilon = a_n$ entails using theorem 4:

$$R_p(\delta, \theta_2) \geqslant \big(L' \rho_n^{s'} g(0) - 2C a_n n^{-\gamma} n^\gamma\big)^p \geqslant \big(L' \rho_n^{s'} g(0)(1 - o_n(1))\big)^p,$$

since $\lim_n a_n/\rho_n^{s'} \to 0$, and the theorem follows. □

Proof of theorem 3. Theorem 3 is an immediate consequence of proposition 1. Clearly, $B \subset A$ thus equations (4.3) and (4.4) entail that $\widetilde{f}_n$ is superefficient at any function $f_0 \in B$. More precisely, $\widetilde{f}_n$ satisfies (7.18) with

$$\gamma = \frac{(s - s')(\beta + 1)}{2(1 + 2s' + \beta)(1 + 2s + \beta)} > 0$$

since $n^{-\gamma}\ell(1/n) \to 0$ for any $\ell \in \mathrm{RV}(0)$. The conclusion follows from proposition 1. □

Appendix A. Some facts on regular variation

We recall here briefly some results about regularly varying functions. The results stated in this section can be found in Bingham et al. (1989). In all the following, let ℓ be a slowly varying function. An important fact is that the property

$$\lim_{h \to 0^+} \ell(yh)/\ell(h) = 1 \tag{A.1}$$

actually holds *uniformly* for y in any compact set of $(0, +\infty)$. If $R_1 \in \mathrm{RV}(\alpha_1)$ and $R_2 \in \mathrm{RV}(\alpha_2)$, we have

$$R_1 \times R_2 \in \mathrm{RV}(\alpha_1 + \alpha_2) \text{ and } R_1 \circ R_2 \in \mathrm{RV}(\alpha_1 \times \alpha_2).$$

If $R \in \mathrm{RV}(\gamma)$ with $\gamma \in \mathbb{R} - \{0\}$, we have

$$R(h) \to \begin{cases} 0 & \text{if } \gamma > 0, \\ +\infty & \text{if } \gamma < 0, \end{cases} \tag{A.2}$$

as $h \to 0^+$. If $\gamma > -1$, we have:

$$\int_0^h t^\gamma \ell(t)dt \sim (1+\gamma)^{-1} h^{1+\gamma} \ell(h) \text{ as } h \to 0^+, \tag{A.3}$$

and $h \mapsto \int_0^h t^\gamma \ell(t)dt$ is regularly varying with index $1+\gamma$. This result is known as the Karamata theorem. Let us define (R is continuous)

$$R^{\leftarrow}(y) = \inf\{h \geqslant 0 \text{ such that } R(h) \geqslant y\},$$

which is the generalized inverse of R. If $R \in \mathrm{RV}(\gamma)$ for some $\gamma > 0$, there exists $R^- \in \mathrm{RV}(1/\gamma)$ such that

$$R(R^-(h)) \sim R^-(R(h)) \sim h \text{ as } h \to 0^+, \tag{A.4}$$

and R^- is unique up to an asymptotic equivalence. Moreover, one version of R^- is $R^{\leftarrow}$.

Acknowledgements. I wish to thank my adviser Marc Hoffmann for helpful advices and encouragements.

References

Antoniadis, A., Gregoire, G. and Vial, P. (1997). Random design wavelet curve smoothing. *Statistics and Probability Letters*, **35** 225–232.

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Statist.*, **6** 127–146 (electronic).

Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1989). *Regular Variation*. Encyclopedia of Mathematics and its Applications, Cambridge University Press.

Brown, L. and Cai, T. (1998). Wavelet shrinkage for nonequispaced samples. *The Annals of Statistics*, **26** 1783–1799.

Brown, L. D. and Low, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimations. *The Annals of Statistics*, **24** 2524–2535.

Buckley, M., Eagleson, G. and Silverman, B. (1988). The estimation of residual variance in nonparametric regression. *Biometrika*, **75** 189–199.

Cai, T. T., Low, M. and Zhao, L. H. (2004). Tradeoffs between global and local risks in nonparametric function estimation. Tech. rep., Wharton, University of Pennsylvania, http://stat.wharton.upenn.edu/~tcai/paper/html/Tradeoff.html.

Delouille, V., Simoens, J. and Von Sachs, R. (2004). Smooth design-adapted wavelets for nonparametric stochastic regression. *Journal of the American Statistical Society*, **99** 643–658.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81** 425–455.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society. Series B. Methodological*, **57** 371–394.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Monographs on Statistics and Applied Probability, Chapman & Hall, London.

Gaïffas, S. (2004). Convergence rates for pointwise curve estimation with a degenerate design. *Mathematical Methods of Statistics*. To appear, available at http://hal.ccsd.cnrs.fr/ccsd-00003086/en/.

Gasser, T., Sroka, L. and Jennen-Steinmetz (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73** 625–633.

Goldenshluger, A. and Nemirovski, A. (1997). On spatially adaptive estimation of nonparametric regression. *Mathematical Methods of Statistics*, **6** 135–170.

Kerkyacharian, G. and Picard, D. (2004). Regression in random design and warped wavelets. *Bernoulli*, **10** 1053–1105.

Lepski, O. V. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and its Applications*, **35** 454–466.

Lepski, O. V., Mammen, E. and Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, **25** 929–947.

Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, **25** 2512–2546.

Maxim, V. (2003). *Restauration de signaux bruités sur des plans d'experience aléatoires*. Ph.D. thesis, Université Joseph Fourier, Grenoble 1.

Spokoiny, V. G. (1998). Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *The Annals of Statistics*, **26** 1356–1378.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, **8** 1348–1360.

Tsybakov, A. (2003). *Introduction l'estimation non-paramtrique*. Springer.

Wong, M.-Y. and Zheng, Z. (2002). Wavelet threshold estimation of a regression function with random design. **80** 256–284.

Laboratoire de Probabilités et Modèles Aléatoires, U.M.R. CNRS 7599 and Université Paris 7, 175 rue du Chevaleret, 75013 Paris

E-mail address: gaiffas@math.jussieu.fr